What is claimed is:~~1. A kit for the production of~~

1. A method of assembling nucleic acid components in a predetermined order to produce a nucleic acid multicomponent ~~constructs, comprising a package containing at least three nucleic acid components, including a first nucleic acid component selected from a category of first nucleic acid components having a common biological utility or functionality, a second nucleic acid component selected from a category of second nucleic acid components having a common biological utility or functionality, and a third nucleic acid component selected from a category of third nucleic acid components having a common biological utility or functionality and, optionally, at least one oligonucleotide bridge,~~

~~wherein the biological utility or functionality of each of the first, second and third categories is different from the other categories,~~

~~wherein at least two of the at least three nucleic acid components are supplied as a library of two or more different nucleic acid fragments, and~~wherein ~~each~~ of the first, second, and third nucleic acid components comprises a double stranded nucleic acid molecule having at least one unique ~~non-palindromic~~**construct, comprising:**

(a) providing at least two nucleic acid components, each comprising at least one genetic element providing a functionality and at least one single stranded 5' or 3' terminal sequence, ~~which allows~~**wherein the terminal sequence hybridizes to either a terminal sequence in another of said at least two nucleic acid components or to an oligonucleotide adaptor molecule that is supplied in addition to the at least two nucleic acid components so as to allow** for specific annealing and linkage of ~~the first, second, and third nucleic acid component, and optionally the at least one oligonucleotide bridge, in a predetermined order.2. The kit~~**all of the nucleic acid components in a predetermined order;**

(b) incubating the nucleic acid components under conditions which allow for specific annealing and assembling of the components to thereby produce the nucleic acid multicomponent construct.

2. The method of claim 1, wherein ~~each~~**one or more** of the ~~first, second, and third nucleic acid components is appropriately phosphorylated for ligation.~~

~~3. A kit for the production of vectors, comprising a package containing at least three nucleic acid components, including an origin of replication, a selectable marker, and an insert of interest, and, optionally, at least one oligonucleotide bridge, wherein at least two of the at least three nucleic acid components are supplied as a library of two or more different nucleic acid fragments, and wherein each of the origin of replication, the selectable marker, and the insert of interest comprises a double standard nucleic acid molecule having at least one unique non-palindromic single stranded 5' or 3' terminal sequence, which allows for specific annealing and linkage of the origin of replication, the selectable marker, and the insert of interest, and optionally the at least one oligonucleotide bridge, in a predetermined order.4. A method of obtaining one or more species of a nucleic acid construct optimal for a particular application comprising:~~

~~(a) providing at least three nucleic acid components, including~~

~~(i) a first component comprising a library of two or more different nucleic acid fragments, each supplying a specific biological utility or functionality chosen from a first category of biological utilities or functionalities,~~

~~(ii) a second component comprising a library of two or more different nucleic acid fragments, each supplying a specific biological utility or functionality chosen from a second category of biological utilities or functionalities, and(iii) a third component comprising one or more third nucleic acid fragments, each supplying a specific biological utility or functionality chosen from a third category of biological utilities or functionalities, wherein said nucleic acid fragments are double stranded nucleic acid molecules having at least one unique nonpalindromic~~**nucleic acid components provide a single functionality.**

3. The method of claim 1, wherein one or more of the nucleic acid components provide multiple functionalities.

4. The method of claim 1, wherein each of the nucleic acid components is flanked by two single stranded 5' or 3' terminal ~~sequence which facilitates specific annealing and linkage of said nucleic acid fragments in a specific order; and~~

~~wherein the biological functionality or utility of each of the first, second, and third categories is different from that of the other categories,~~

~~(b) contacting said first, second and third nucleic acid components so that said nucleic acid fragments become specifically annealed and linked so as to produce a library of at least 4 different vector sequences from which one or more species optimal for a particular application may be obtained, and(c) isolating one or more species optimal for a particular application~~**sequences**.

5. ~~A method for generating a library of vectors, comprising(i) providing at least 3 different types of nucleic acid components, each representing a functionality to be included in a vector, wherein~~

~~(a) at least two of said nucleic acid components types are provided as a library of two or more different nucleic acids, and(b) said nucleic acid components are double-stranded nucleic acid molecules having at least one unique non-palindromic single stranded 5' or 3' terminal sequence which facilitates specific annealing and linkage of said components; and~~

~~(ii) annealing and ligating said nucleic acid components to generate a library of at least 4 different vector sequences.6. The method of claim 4 or 5, wherein a library of at least 27 different vector sequences is generated.7. The method of claim 6, wherein a library of at least 625 different vector sequences is generated.8. The method of claim 7, wherein a library of at least 1024 different vector sequences is generated.~~**The method of claim 1, wherein at least one of the single stranded terminal sequences are non-palindromic.**

6. The method of claim 1, wherein the nucleic acid components are incubated

simultaneously.

7. The method of claim 1, wherein the nucleic acid components are incubated in a step-wise fashion.

8. The method of claim 1, wherein the nucleic acid components are linked directly via annealing of 5' complementary terminal sequences.

9. The method of claim ~~4 or 5, wherein said specific biological utility or functionality is selected from the group consisting of: an origin of replication, a selectable marker gene, a protein coding region, a transcription regulatory element, and a translation regulatory element.10. The method of claim 9, wherein the origin of replication is selected from the group consisting of: a bacterial origin of replication, a viral origin of origin of replication, a phage origin of replication, a eukaryotic origin or replication, a yeast origin of replication, and a mammalian origin of replication.11. The method of claim 9, wherein the selectable marker gene is selected from the group consisting of: an antibiotic resistance selectable marker, a drug resistance selectable marker, and a mutagenic resistance selectable marker.12. The method of claim 5, wherein the library of vectors is selected from the group consisting of a library of DNA virus vectors, a library of RNA virus vectors, a library of single stranded virus vectors, a library of double stranded virus vectors, a library of baculovirus vectors, a library of retrovirus vectors, a library of adenovirus vectors, a library of adeno associated virus vectors, a library of Herpes virus vectors, a library of Vaccinia virus vectors, a library of replication defective retrovirus vectors, and a library of artificial chromosome vectors.13. The method of claim 4, wherein the one or more optimal species isolated is optimized for a function selected from the group consisting of: optimal expression, optimal gene transfer activity, and optimal gene therapy activity.14. The kit of claim 1, wherein at least two of said at least three nucleic acid components are supplied as at least three different nucleic acid species, each of which provides an alternative form of the common biological utility or functionality of the nucleic acid component.15. The kit~~**1, wherein the nucleic acid components are linked directly via annealing of 3' complementary terminal sequences.**

10. The method of claim 1, wherein the nucleic acid components are linked indirectly via a linking nucleic acid molecule, the linking nucleic acid molecule comprising an oligonucleotide.

11. The method of claim 1, wherein at least two of the nucleic acid components are linked indirectly via an adaptor molecule which provides terminal sequences that are complementary with 5' or 3' terminal sequences of the at least two nucleic acid components.

12. The method of claim 1, wherein the unique single stranded, non-palindromic terminal sequences have a length of 10 bases.

13. The method of claim 1, wherein the unique single stranded, non-palindromic terminal sequences have a length of 20 bases.

14. The method of claim 1, wherein steps (a) and (b) are repeated with one or more of the nucleic acid components substituted with a different nucleic acid component chosen from a category of components, having the same functionality or characteristic utility, but possessing the same terminal sequences which allow for linkage and production of a different nucleic acid construct.

15. The method of claim 1, wherein the ~~common~~**nucleic acid component encodes a** biological ~~utility or~~ functionality ~~is~~ selected from the group consisting of~~:~~ ~~an~~ origin of replication, ~~a~~ selectable marker, ~~a promoter, a regulatory element, a translational element, a transcriptional terminator, a sequence which regulates mRNA stability, a subcellular localization element, a recombinational element, a~~**transcriptional regulatory element, structural gene or fragment thereof, transcription termination signal, translational regulatory sequence, regulators of mRNA stability, cellular localization signal, recombination elements, mutagenized genes, protein domain encoded regions,** synthetic multiple cloning ~~site, a~~**sites,** unique restriction enzyme or DNA cleavage ~~site, a~~**sites, and** site for ~~the~~ covalent or ~~noncovalent~~**non covalent** attachment of a biological or chemical molecule~~, and a protein coding sequence~~.

16. The ~~kit~~**method** of claim 15, wherein the ~~regulatory element is selected from the group consisting of: a promoter, and enhancer, a polyadenylation signal.17. The kit of claim 15, wherein the protein coding sequence~~**DNA cleavage site is part of a multiple cloning site.**

17. The method of claim 1, wherein the nucleic acid component is covalently or non-covalently modified.

18. The method of claim 17, wherein the modification is biotinylation.

19. The method of claim 17, wherein the modification is fluorescent tagging.

20. The method of claim 17, wherein the modification is incorporation of polypeptide nucleic acids (PNA).

21. The method of claim 17, wherein the modification is covalent or non-covalent conjugation of a protein involved in nucleic acid modification.

22. The method of claim 21, wherein the protein involved in nucleic acid modification is an enzyme.

23. The method of claim 17, wherein the modification is covalent or non-covalent conjugation of a protein or another molecule or ion which enables the recognition and binding of a specific molecular target.

24. The method of claim 23, wherein the specific molecular target is a hapten.

25. The method of claim 1, wherein annealing and linkage of step (b) is achieved by heating, followed by cooling down to an appropriate temperature, such that efficient annealing of the nucleic acid component terminal sequences occurs.

26. The method of claim 25, further comprising treating with T4 DNA ligase to ligate the nucleic acid components.

27. The method of claim 1, wherein the nucleic acid construct is selected from the group consisting of~~: a cDNA, a structural gene, a fragment of a structural gene, a mutagenized structural gene, and a mutagenized fragment of a structural gene.~~ **a vector, a cDNA library, a phage or viral genome, and a gene or gene fragment.**

28. The method of claim 27, wherein the gene is a mutagenized gene.

29. The method of claim 27, wherein the gene is a combined fusion gene.

30. The method of claim 27, wherein the gene is an artificial gene.

31. A method of producing a vector, comprising:

a) providing at least two nucleic acid components and optionally a linking nucleic acid molecule to be assembled into the construct, each component comprising a double stranded nucleic acid molecule encoding a functionality and having at least one single stranded 5' or 3' terminal sequence, wherein the terminal sequence hybridizes to either a terminal sequence in a separate nucleic acid component or to a sequence in a linking nucleic acid molecule so as to allow for specific annealing and linkage of the components in a predetermined order, wherein the nucleic acid components encode:

i) an origin of replication

ii) a selectable marker

iii) an insert of interest;

(b) incubating the nucleic acid components under conditions which allow for specific annealing and linkage of the nucleic acid components to thereby produce the functional vector.

32. The method of claim 31, for producing a cosmid vector, further comprising providing a nucleic acid component encoding a lambda phage cohesive end (cos site).

33. The method of claim 31, for producing a lambda phage vector, further comprising providing nucleic acid components encoding a left and a right arm of the lambda phage genome.

34. The method of claim 31, for producing a retroviral vector, further comprising providing a nucleic acid component encoding a retroviral genome including long terminal repeats (LTR).

35. The method of claim 31, for producing a yeast artificial chromosome, further comprising providing nucleic acid components encoding a yeast centromere and two yeast telomeres.

36. The method of claim 31, for producing a vector expressing a protein of interest, further comprising providing a nucleic acid component encoding a structural gene of interest.

37. The method of claim 31, for producing a vector expressing a cDNA library further comprising, providing nucleic acid components encoding a collection of cDNA molecules derived from poly(A)+ mRNA.

38. The method of claim 31, for producing a vector expressing a genomic library, further comprising providing nucleic acid components encoding a collection of gene or gene fragments derived from the genome of an organism.

39. A kit for the production of nucleic acid multicomponent constructs, comprising a package containing nucleic acid components, each component comprising a double stranded nucleic acid molecule encoding a functionality and having at least one single stranded 5' or 3' terminal sequence, wherein the terminal sequence hybridizes to either a terminal sequence in a separate nucleic acid component or to a sequence in a linking nucleic acid molecule so as to allow for specific annealing and linking of the components in a predetermined order.

40. A kit for the production of nucleic acid multicomponent constructs, comprising at least 3 different nucleic acid components, each encoding a functionality, and including a 5' OH terminal phosphate group for ligation, the kit further comprising a ligase enzyme.

41. A kit for the production of vectors, comprising nucleic acid components, each component comprising a double stranded nucleic acid molecule encoding a functionality and having at least one single stranded 5' or 3' terminal sequence, wherein the terminal sequence hybridizes to either a terminal sequence in a separate nucleic acid component or to a sequence in a linking nucleic acid molecule so as to allow for specific annealing and assembling of the components in a predetermined order, wherein the nucleic acid components encode: i) an origin of replication, and ii) a selectable marker.

42. A method of assembling nucleic acid components in a predetermined order to produce a nucleic acid multicomponent construct, comprising:

(a) providing the nucleic acid components and one or more linking nucleic acid molecules into the construct, each nucleic acid component comprising a double stranded nucleic acid molecule encoding a functionality having at least one single stranded 5' or 3' terminal sequence, wherein the terminal sequence hybridizes to a sequence in a linking nucleic acid molecule so as to allow for specific annealing of complementary sequences and linkage of the components in a predetermined order;

(b) incubating the nucleic acid components which allow for the specific annealing and

assembling of the nucleic acid components to thereby produce the nucleic acid multicomponent construct.

43. A method for generating a nucleic acid construct by assembling nucleic acid components in a predetermined order, comprising:

(a) providing at least three nucleic acid components, each independently comprising a nucleic acid sequence having a biological functionality or characteristic utility, and optionally, at least one oligonucleotide bridge, wherein each of said nucleic acid components are double stranded nucleic acid molecules having at least one unique nonpalindromic single stranded 5' or 3' terminal sequence which facilitates specific hybridization of said nucleic acid components, and said oligonucleotide bridge(s) if provided, to assemble said nucleic acid components in a predetermined order; and

contacting said at least three nucleic acid components, and said oligonucleotide bridge(s) if provided, under conditions in which said 5' and 3' terminal sequences hybridize to generate a multicomponent nucleic acid construct.

BACKGROUND OF THE INVENTION

The essence of recombinant DNA technology is the joining of two or more separate segments of DNA to generate a single DNA molecule that is capable of autonomous replication in a given host. The simplest constructions of hybrid DNA molecules involve the cloning of a DNA sequence of interest (**such as** DNA insert **containing a natural or synthetic gene or gene fragment**) into a pre-assembled cloning vector. The cloning vector includes all of the necessary components for replication of the DNA insert in a compatible host cell, e.g., promoter sequence, origin of replication sequence, termination sequence, and a selectable marker sequence. The DNA insert sequences can be derived from essentially any organism, and they may be isolated directly from the genome, from mRNA, or from previously cloned DNA sequences. Alternatively, the DNA insert sequences can be created synthetically.

Insertion of the DNA sequence of interest can be accomplished by a number of techniques. The most common technique involves restriction enzymes. A restriction enzyme recognition site that is present in both the DNA insert and the vector of interest is cleaved with a restriction enzyme to provide for appropriate termini, the termini of either the DNA insert or the vector are treated with alkaline phosphatase to remove terminal phosphates and avoid undesirable joining, and the DNA sequence of interest is inserted into the vector at the compatible sites during a ligation reaction. A restriction enzyme site present in a pre-assembled vector must be compatible with a restriction enzyme site in the DNA sequence of interest.

Alternatively, the DNA of interest can be modified to obtain compatible restriction sites by filling in of cohesive ends as appropriate, or by the ligation of an appropriate oligonucleotide linker, which can be subsequently cleaved by the restriction enzyme of interest.

Conventional cloning methods can be time consuming and often involve multiple sub cloning steps. Therefore, a need exists for developing a simple and rapid method for synthesizing and

identifying an optimal construct for use in a particular application.

SUMMARY OF THE INVENTION

This invention pertains to methods for preparing multicomponent nucleic acid constructs. The ~~invention~~**method of the invention has a wide variety of applications for the expression of synthetic and naturally occurring genes or gene fragments. The invention provides integral vector elements which may be specifically selected or which may varied so as to create a collection of vectors from which optimal configurations can be selected or screened for. These integral vector elements include both vector backbone elements, which do not directly affect the expression or form of the insert gene or gene fragment, and insert modifying vector elements which alter the expression and/or form of the insert-encoded gene product. This system for the rapid and flexible assembly of specific multicomponent nucleic acid constructs is referred to as GEOS (for Genetic Engineering Operating System). GEOS methodology has numerous applications ranging from the assembly of simple circular expression vectors to the production of complex linar assemblies which function as small chromosomes. Various applications of GEOS methodology are discussed in detail below and still others will be apparent to the skilled artisan.**

The invention further provides a method of linking **the vector element** nucleic acid components in a predetermined order **so as** to produce a nucleic acid multicomponent construct~~, comprising:~~**.**

In certain preferred embodiments, the GEOS method comprises:

(a) providing the nucleic acid components ~~and optionally a linking nucleic acid molecule~~ to be assembled into the construct, each nucleic acid component comprising a double stranded nucleic acid molecule having at least one single stranded 5' or 3' terminal sequence, the terminal sequence having sufficient complementarity to ~~either~~ a terminal sequence in a separate nucleic acid component ~~or to a sequence in a linking nucleic acid molecule~~ so as to allow for specific annealing of complementary sequences and linkage of the components in a predetermined order;

(b) incubating the nucleic acid components under conditions which allow for the specific annealing and linkage of the nucleic acid components to thereby produce the nucleic acid multicomponent construct.

In another preferred embodiments, the GEOS method comprises:

(a) providing the nucleic acid components and one or more linking nucleic acid molecules to be assembled into the construct, each nucleic acid component comprising a double stranded nucleic acid molecule having at least one single stranded 5' or 3' terminal sequence, the terminal sequence having sufficient complementarity to a sequence in a linking nucleic acid molecule so as to allow for specific annealing of complementary sequences and linkage of the components in a predetermined order;

(b) incubating the nucleic acid components under conditions which allow for the specific

annealing and linkage of the nucleic acid components to thereby produce the nucleic acid multicomponent construct.

The genetic element portion(s) of the nucleic acid components can be double- or single-stranded, though are preferably double-stranded.

In a preferred embodiment of the method, the nucleic acid components are flanked by single stranded terminal sequences and these terminal sequences are preferably non-palindromic. The nucleic acid components can be linked either directly via annealing of 5' or 3' complementary terminal sequences or indirectly via a linking nucleic acid molecule (e.g. an oligonucleotide or an adaptor molecule).

The nucleic acid components can be linked either simultaneously or sequentially to form the nucleic acid construct. Sequential assembly is suitable for automation. The method can be used to produce nucleic- acid constructs which are functional as assembled or constructs which are used as subcomponents for the assembly of functional constructs.

The method of the invention can be used to synthesize a group of nucleic acid constructs in which one or more of the components can be substituted, in each of the constructs, with a different nucleic acid component, having the same functionality or characteristic utility. This allows for comparison of the different components and production of an optimal construct for a particular application. Toward this end, the nucleic acid components are designed and synthesized in such a way that a group of nucleic acid components belonging in the same category (i.e., having the same functionality or characteristic utility, e.g. a set of nucleic acid components encoding different promoters) possess the same terminal sequences, such that the same category nucleic acid components can be used interchangeably to assemble a nucleic acid multicomponent construct.

The nucleic acid components may also be covalently or non-covalently modified prior to or following assembly of the nucleic acid multicomponent construct. This allows for the synthesis of constructs having biological properties which cannot be obtained easily using current recombinant methods. **For instance, the modification utilizes an arylboronic acid reagent, such as phenyldiboronic acid.**

The method of this invention is particularly suitable for the construction of nucleic acid vectors. These include plasmid, viral, or phage vectors, or yeast artificial chromosomes. The vector can be a cloning or expression vector and can be used for the expression of cDNA or genomic libraries, genes or gene fragments, mutagenized genes, recombined fusion genes, and artificial genes. The constructs can be employed in prokaryotic, eukaryotic (mammalian or non-mammalian) expression, construction of unique cDNA libraries, protein, antibody and peptide phage display libraries. The constructs can further be employed in gene transfer, gene therapy, and the creation of transgenic organisms.

According to the method, the vector is assembled from nucleic acid components encoding a single functionality or multiple functionalities. ~~At a minimum~~**In some applications of the invention, more than one biological function may be bundled into a single nucleic acid**

component. This may be desirable when, for example, one seeks to limit the overall number of components to be assembled into the GEOS construct. In one embodiment of the invention, nucleic acid components encoding an origin of replication, a selectable marker and an insert of interest are used. Depending on the type of vector desired, nucleic acid components encoding other vector functions may also be incorporated (e.g. a promoter, a transcription or translation regulatory element, etc.). An expression vector can be produced using a nucleic acid component encoding a structural gene or gene fragment of interest and additional nucleic acid components encoding regulatory elements required for expression of the gene. For example, a cDNA library expression vector is produced using nucleic acid components encoding a collection of cDNA molecules derived from poly(A)+ mRNA. Importantly, the optimization procedure of interchanging nucleic acid components described above can be used to create an optimal vector for a particular application.~~The reagents required to practice the method of the invention may be provided in the form of a kit. A kit would comprise, in separate containers, the nucleic acid components to be assembled into a construct, and optionally linking nucleic acid molecules as well as buffers, enzymes and an instructional brochure explaining how to use the kit. In a preferred embodiment the kit would provide the nucleic acid components in an appropriately phosphorylated form for ligation.~~

The invention further provides a kit for the production of vectors. In one embodiments, the kit for the production of vectors would minimally comprise nucleic acid components encoding origins of replication and selectable markers and optionally, transcriptional regulatory sequence(s). The kit could also include nucleic acid components encoding other vector functions (e.g. a promoter, a transcription or translation regulatory element, etc.).

The invention further provides a kit for the production of vectors. The kit for the production of vectors would minimally comprise nucleic acid components encoding origins of replication, selectable markers and inserts of interest. The kit could also include nucleic acid components encoding other vector functions (e.g. a promoter, a transcription or translation regulatory element, etc.).

The method of the invention is a highly efficient, rapid, cost effective alternative to current recombinant cloning methods in that it enables users to choose from a broad array of different nucleic acid components or modified nucleic acid components when assembling any construct. The method of the invention allows the rapid construction of customized constructs without the need to use restriction enzymes.

Other features and advantages of the invention will be apparent from the following detailed description, and from the claims.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 is a schematic representation of the assembly of a circular plasmid, using the method of the invention. The plasmid vector is assembled by combining a set of nucleic acid components which possess complementary terminal sequences, as well as all of the necessary genetic elements required to generate a functional plasmid construct. A partial list of different interchangeable nucleic acid components and their respective categories is shown, demonstrating

the flexibility and utility of the method of the invention.

FIG. 2 shows representative ways of linking nucleic acid components via specific terminal sequences to prepare nucleic acid constructs according to the method of the invention.

FIG. 2(A) shows annealing of non-palindromic complementary terminal sequences;

FIG. 2(B) shows annealing of 5' compatible terminal sequences;

FIG. 2(C) shows annealing of 3' compatible terminal sequences;

FIG. 2(D) shows linking of non-compatible terminal sequences via an oligonucleotide bridge (thick line);

FIG. 2(E) shows linking of non-compatible terminal sequences via an adaptor (thick lines).

~~DETAILED DESCRIPTION~~

FIG. 3 illustrates a method for linking nucleic acid components via a specific semi-cruciform oligonucleotide bridge (thick line) which itself is composed of two partially complementary oligonucleotides. The two nucleic acid component pieces, one with a 5' overhang and the other with a 3' overhang, are brought together by the semi-cruciform bridge and subsequently covalently joined by a ligation step as shown.

FIG. 4 illustrates a method for introducing unique 3' overhangs at each end of a vector element by a method employing pcr amplification with primers containing a phosphorothioate-nucleotide linkage followed by exonuclease digestion of the amplification product to create unique complementary 3' overhangs in the vector elements to be conjoined.

FIG. 5 is a schematic representation of a illustrative method for carrying out the subject GEOS combinatorial method in a manner that utilizes flanking intronic sequences to generate a combinatorial gene library.

DETAILED DESCRIPTION OF THE INVENTION

I. Overview

One of the most powerful techniques in molecular biology involves the use of restriction enzymes for the purpose of cloning DNA inserts into a specific site in a cloning vector. However, as described herein, the use of restriction-based cloning techniques has certain limitations, particularly with regard to generation of multicomponent combinatorial libraries. The invention described herein enables the rapid and precise generation of a wide range of nucleic acid constructs, including highly optimized vectors containing cloned inserts, without the use of restriction enzymes and without prior knowledge of the sequence of the cloned insert.

The present method is based on the ability to assemble a multitude of individual nucleic acid components, including genes, gene fragments and other genetic elements, into a useful nucleic acid construct such as a vector. The invention further provides a method for linking nucleic acid components in a predetermined order to produce a nucleic acid construct. Components incorporate unique and specifiable terminal sequences, or overhangs, which are preferably non-palindromic and single stranded, and which serve to direct the site-specific localization of each component within an assembled construct, such that one overhang, A, will anneal preferably with a discrete, complementary overhang, A', located on an adjacent component. Several altenative methods for joining adjacent components are described in FIG. 2, including those that might also utilize oligonucleotide bridges and adaptor molecules.

Categories of components having the same functionality, or characteristic utility, can be designed and synthesized to incorporate the same overhangs, thereby enabling the interchangeable deployment of components from a specific category. According to the method, the construct is assembled from nucleic acid components encoding a single functionality or multiple functionalities. In some applications of the invention, more than one functionality may be bundled into a single nucleic acid component especially when, for example, one wishes to limit the overall number of components to be assembled into the construct. As described herein, the present invention provides a method for creating combinatorial arrays of constructs from a limited set of nucleic acid components.

In one embodiment of this method, the generation of a wide array of candidate DNA vaccine constructs is greatly simplified. When specifically applied to somatic gene immunization, the current method enables the rapid assembly and systematic variation of critical components, including a general component for the expression vector backbone, components corresponding to framework regions, components corresponding to complementarity-determining regions (CDRs), components corresponding to tissue-specific immunoglobulin promoter and/or enhancer elements, and one or more antigenic epitope components to be inserted into any one or more of the CDR domains.

Novel genes and gene products can also be generated utilizing this invention. In one embodiment, a library of antibodies can be displayed on the surface of a lambda phage by selecting individual components from a category of cDNA components representative of an antibody framework region, a category of cDNA components representative of a hypervariable region, and one or more user-definable components which incorporate user-specifiable genetic elements required to express antibody fusions on the surface of the phage.

Other aspects of the invention show that components may be covalently modified or altered prior to assembly, and following assembly, constructs would incorporate one or more of these modifications. These modifications may act as the site of attachment for small biological molecules or macromolecular biological molecules, including proteins and carbohydrates. The ease and flexibility of modification to components rather than constructs may improve the specificity of many gene therapy vectors, especially those

requiring import into the nucleus in order for expression to occur. In one embodiment, the design of candidate vector therapeutic constructs would involve the deployment of a multitude of user-definable vector components, resulting in the generation of a combinatorial array of vectors. A discrete category of components to facilitate nuclear targeting, able to be covalently modified using an array of cationic peptides, could then be assembled with other components into an exquisitely specifiable construct.

In another embodiment, covalent modification of a component using methods know in the art can result in the reversible but stable attachment of that component to a solid phase. In such an example, the subsequent stepwise addition of components containing compatible overhangs would result in a uniquely automatable process for assembling arrays of vectors or other constructs.

In another embodiment, both viral and phage genomes, as well as vector constructs, may utilize altered or mutagenized components. Such components would enable the mutation or deletion of one or more genes, the enhancement of specific gene functions, the construction of fusion genes, or, for instance, the addition or deletion of restriction enzyme sites.

In a particularly preferred embodiment of the invention, individual components comprise exonic and intronic units which can used to facilitate rearrangement of discrete polypeptide-encoding exon units. The invention thereby facilitates the formation of peptide domain "shuffled" libraries encoded by exonic units which are linked by intronic units. Thus the invention can be readily applied to the creation of unique proteins peptide domain shuffling and in vitro selection techniques. In one particular embodiment of the peptide domain shuffling application a pool of diverse peptide domain-encoding nucleic acid components is added to a construct assemblage. A compatible intronic unit is then added prior to the addition of another pool of diverse peptide domain-encoding exonic nucleic acid components. The resulting assembly increases in heterogeneity with successive addition of peptide domain-encoding components and provides a convenient source of diversity for subsequent screening or selection processes. This methodology thereby enables the user to create synthetically "evolved" proteins comprised of discrete polypeptide domains which have been randomized and infinitely "shuffled."

The invention further provides a kit for the production of vectors. In one embodiments, the kit for the production of vectors would minimally comprise nucleic acid components encoding origins of replication and selectable markers and optionally, transcriptional regulatory sequence(s). The kit could also include nucleic acid components encoding other vector functions (e.g. a promoter, a transcription or translation regulatory element, etc.).

The invention further provides a kit for the production of vectors. The kit for the production of vectors would minimally comprise nucleic acid components encoding origins of replication, selectable markers and inserts of interest. The kit could also include nucleic acid components encoding other vector functions (e.g. a promoter, a transcription or translation regulatory element, etc.).

II. Definitions

In order that the invention may be more readily understood, certain terms are first defined.

As used herein, the term "~~nucleic acid component" describes~~**GEOS" stands for Genetic Engineering Operating System and is meant to refer to the generalized method of the immediate invention which employs both a flexible strategy for vector assembly from selected nucleic acid components encoding various biological functions, and which further employs methods for the rapid chemical joining of these selected nucleic acid components to create expression vectors which are ideally suited to a particular purpose.**

As used herein, the term "vector" is intended to include both circular and linear assemblies of nucleic acid components. Examples of linear vectors include various viral genomes as well as yeast artificial chromosomes (YACs) and mammalian artificial chromosomes (see e.g. Grimes and Cooke (1998) Hum Mol Genet, 7: 1635-40; Vos (1998) Curr Opin Genet Dev, 8: 351-9).

The terms "nucleic acid component", "vector component" and "vector element", which used interchangeably herein, describe the basic unit of assembly used in the present invention. ~~Nucleic acid components~~**These units** are comprised of **nucleic acid molecules, preferably** double stranded ~~nucleic acid molecules,~~ which contain at their termini specific terminal sequences required for assembling the nucleic acid components into a specific nucleic acid multicomponent construct. The nucleic acid sequences contained within each nucleic acid component provide the requisite information for a specific biological function or **functions or** for a specific utility deemed essential by the user. Examples of nucleic acid components include ~~the~~ nucleic acid sequences which encode a ~~gene.~~**polypeptide, include** an origin of replication, ~~or a selection marker.~~**and/or include a selectable marker, alone or in combination with other biologically active nucleotide sequences.**

The term "nucleic acid" refers to polynucleotides such as deoxyribonucleic acid (DNA), and, where appropriate, ribonucleic acid (RNA). The term should also be understood to include, as equivalents, analogs of either RNA or DNA.

As used herein, the term "terminal sequence" is used to describe the terminal single stranded nucleotide sequence of a nucleic acid component. Nucleic acid components having complementary terminal sequences to either separate nucleic acid components or linking molecules enable users to specify the precise organization and orientation of nucleic acid components upon their assembly into constructs.

The terms "complementary" and "compatible" are used herein interchangeably to describe the capacity of a pair of single-stranded terminal sequences to anneal to each other via base pairing (e.g. A-T or G-C). The terminal sequences should contain nucleotide sequences of sufficient length and sequence complementarity so as to allow efficient annealing to occur.

As used herein, the term "palindromic sequence"**, which is art recognized,** describes a sequence of DNA that consists of inverted repeats.

As used herein, the term "linkage" refers to a physical connection**, preferably covalent coupling,** between two or more nucleic acid components, **e.g.,** catalyzed by an enzyme~~,~~ **ushc as a ligase.**

As used herein, the term "genomic library" refers to a set of cloned fragments together representing the entire genome of an organism.

As used herein, the term "category" describes a classification of genes, gene fragments, restriction sites, or **other** genetic elements **found in the subject nucleic acid components** which may be arranged in a systematic order based on a number of user-defined criteria, including the ability to produce or regulate a similar biological activity. For example, the various different origin of replication nucleotide sequences, may be classified ~~into~~**as** a specific category. **Marker genes, transcriptional regulatory sequence and the like are each other examples of categories of functionality which be provide in the nucleic acid components.**

As used herein, the term "hapten" refers to a small molecule that acts as an antigen when conjugated to a protein.

As used herein, the term "genetic element" describes a sequence of nucleotides, including those which encode a regulatory region, involved in modulating or producing biological activity or responses or which provides a specific signal involved in a molecular mechanism or biological activity. For example, a prokaryotic gene may be comprised of several genetic elements, including a promoter, a protein coding region, a Shine-Delgarno sequence, and translational and transcriptional terminators.

As used herein, the term "functionality" describes the normal, characteristic utility **or utilities** of a construct, gene, gene fragment, or genetic element.

As used herein, the term "handle" is used to describe a chemical or biochemical modification to a nucleotide residue within an oligonucleotide or a nucleic acid component. A handle provides a site for covalent or non-covalent attachment of a biological or chemical molecule(s) to a nucleic acid component.

As used herein, the term "oligonucleotide" refers to a single stranded nucleic acid sequence composed of two or more nucleotides. An oligonucleotide can be derived from natural sources, but it is often chemically synthesized by known methods and then purified. It may be of any length and it may be used as a primer, a probe or a component of a ligation reaction.

As used herein, the term "oligonucleotide bridge" is an oligonucleotide used in a ligation reaction to bridge non **-**complementary 5' and 3' terminal sequences in two separate nucleic acid components.

As used herein, the term "semi-cruciform" refers to a pair of partially complementary oligonucleotides which, when annealed together, function as an oligonucleotide bridge to bring together non-complementary 5' and 3' terminal sequences in two separate nucleic acid components. The two oligonucleotides comprising the bridge include one which

carries, at its 5' end, a sequence which is complementary to the 5' overhang of one of the nucleic acid components and another oligonucleotide which carries, at its 3' end, a sequence which is complementary to the 3' overhand of the other nucleic acid component.

As used herein, the term "promoter" refers to a DNA sequence which is recognized by an RNA polymerase and which directs initiation of transcription at a nearby downstream site. As used herein "promoter" refers to viral, prokaryotic or eukaryotic transcriptional control sequences.

As used herein, the term "enhancer" refers to a DNA sequence which, without regard to its position or its orientation in the DNA, increases the amount of RNA synthesized from an associated promoter. Enhancers are typically found in association with eukaryotic or viral promoters and frequently confer tissue-specific and/or developmental-specific expression of the linked promoter.

As used herein, the term "silencer" refers to a DNA sequence which, without regard to its position or its orientation in the DNA, decreases the amount of RNA synthesized from an associated promoter. Silencers are typically found in association with eukaryotic promoters and frequently confer tissue-specific and/or developmental-specific expression of the linked promoter.

As used herein, the term "transcriptional terminator" refers to a DNA sequence which promotes the formation of a 3' end of an RNA transcript. As used herein the term "transcriptional terminator" refers to viral, prokaryotic or eukaryotic transcriptional terminator sequences (e.g. polyadenylation signal sequences).

As used herein, the term "origin of replication" refers to a DNA sequence which promotes the initiation of DNA synthesis by a DNA polymerase. As used herein the term refers to viral, prokaryotic or a eukaryotic replication origins.

As used herein, the term "exon" refers to a segment of DNA which encodes a portion of a mature RNA transcript.

As used herein, the term "intron" refers to a segment of DNA which encodes a portion of a primary transcript that is not included in a mature RNA transcript. As used herein, the term intron is interchangeable with the term "intervening sequence" and generally refers to a portion of a primary transcript, or the corresponding segment of DNA encoding such a portion, which is removed from a mature RNA transcript by splicing processes.

III. Exemplary GEOS Methodology

The GEOS methodology can be divided into two phases. In one phase, the nucleic acid components are selected and in another phase the selected nucleic acid components are chemically joined. These phases are considered separately below, although it will be obvious to a skilled individual that the two phases are interdependent in that selection of particular nucleic acid components will influence the selection of chemical joining methods.

We begin with an examination of the various chemical joining methods which can be used in the method of the invention.

A. GEOS Chemical Joining Methods

The present invention pertains to a highly efficient, rapid, and cost effective method of producing multicomponent nucleic acid constructs. ~~The~~ **by means of facile chemical joining techniques. In certain preferred embodiments, the GEOS** method comprises:

(a) providing the nucleic acid components ~~and optionally a linking nucleic acid molecule~~ to be assembled into the construct, each nucleic acid component comprising a double stranded nucleic acid molecule having at least one single stranded 5' or 3' terminal sequence, the terminal sequence having sufficient complementarity to ~~either a terminal sequence in a separate nucleic acid component or to a sequence in a linking nucleic acid molecule so as to allow for specific annealing of complementary sequences and linkage of the components in a predetermined order;~~

~~(b) incubating the nucleic acid components under conditions which allow for the specific annealing and linkage of the nucleic acid components to thereby produce the nucleic acid multicomponent construct.~~**a terminal sequence in a separate nucleic acid component so as to allow for specific annealing of complementary sequences and linkage of the components in a predetermined order;**

(b) incubating the nucleic acid components under conditions which allow for the specific annealing and linkage of the nucleic acid components to thereby produce the nucleic acid multicomponent construct.

In another preferred embodiments, the GEOS method comprises:

(a) providing the nucleic acid components and one or more linking nucleic acid molecules to be assembled into the construct, each nucleic acid component comprising a double stranded nucleic acid molecule having at least one single stranded 5' or 3' terminal sequence, the terminal sequence having sufficient complementarity to a sequence in a linking nucleic acid molecule so as to allow for specific annealing of complementary sequences and linkage of the components in a predetermined order;

(b) incubating the nucleic acid components under conditions which allow for the specific annealing and linkage of the nucleic acid components to thereby produce the nucleic acid multicomponent construct.

In many aspects of its practice, the subject method will be carried out in a combinatorial fashion, e.g., to produce a variegated library of multicomponent constructs. In this regard, if there are N component positions in the final multicomponent construct, and Y,sub,N different nucleic acid components at each position N, then the combinatorial library will include {Y,sub,1,times,Y,sub,2,times. . . . Y,sub,N-1,times,Y,sub,N } different multicomponent constructs. To further illustrate, for a three component construct, e.g., N=3, if there are 2 different choices of components at two positions and 4 different choices

of components at the third position, the combinatorial library can include up to 16 different three component constructs (2.times.2.times.4). This point is illustrated further in connection with Table I. In preferred embodiments, N is 3 or greater, more preferably in the range of 3-10, and even more preferably 3-8. Preferably the library includes multicomponent constructs having at least3 different nucleic acid components, rather than being oligomers of a single component.

When the interactions of the nucleic acid components are random, the order and composition of the resulting constructs of the combinatorial library generated is also random. For instance, where the variegated population of nucleic acid components used to generate the combinatorial genes comprises X different components, random assembly of the components can result in X.sup.N different genes having N component positions. Where 5 different nucleic acid components are used (X=5), the combinatorial approach can give rise to 625 different genes having 4 component positions, and 780 different genes having from 1 to 4 component positions (e.g. from the binomial ##EQU1##

or X.sup.1 +X.sup.2 . . . +X.sup.N-1 +X.sup.N). It will be appreciated that the frequency of occurrence of a particular component in the combinatorial library may also be influenced by, for example, varying the concentration of that component positions relative to the other component positions present, or altering the flanking overhang sequences of that component to either diminish or enhance its annealing ability relative to the other component positions being admixed.

However, the present GEOS method can also be utilized for ordered gene assembly, and carried out in much the same fashion as automated oligoucleotide or polypeptide synthesis, such as through the use of resin-bound nucleic acid components in the ordered synthesis of a gene.

In a preferred embodiment of the invention, the nucleic acid components are used in an appropriately phosphorylated form for ligation. Typically, the nucleic acid components are incubated at a temperature appropriate to promote denaturation, cooled down to an appropriate temperature, such that efficient annealing of the nucleic acid component terminal sequences occurs, and treated with a ligase enzyme to ligate the nucleic acid components and produce a nucleic acid construct. The formed nucleic acid construct can be transformed into a bacterial host for amplification and subsequent purification.

The method of the present invention entails the use of specially designed nucleic acid components to assemble a nucleic acid construct. ~~The~~**In one embodiment, the** nucleic acid components are double stranded nucleic acid molecules having one or more, preferably two terminal sequences designed to be complementary to the terminal sequences of the nucleic acid component intended to be the adjacent component in the construct. For example, in a construct containing five components in order 1-5 (see FIG. 1), the terminal sequence E of nucleic acid component 1 would be compatible only with the terminal sequence E', of nucleic acid component 2, the terminal sequence D of nucleic acid component 2 with the terminal sequence D' of nucleic acid component 3, the terminal sequence C of nucleic acid component 3 with the terminal sequence C' of nucleic acid component 4 and the like.

In a preferred embodiment of the method, the nucleic acid components are flanked by single stranded terminal sequences and ~~these~~**the** terminal sequences **of the component** are non-palindromic.

The nucleic acid components can be linked either directly via annealing of 5' or 3' complementary terminal sequences or indirectly via a linking nucleic acid molecule, which can be, for example, a) an oligonucleotide bridge having a sequence that is complementary to 5' and 3' terminal sequences in two separate nucleic acid components or ~~(~~b) an adaptor molecule having terminal sequences that are complementary with 5' or 3' terminal sequences in separate nucleic acid components.

Alternatively, the nucleic acid components may be provided in the form of single stranded nucleic acid molecules, which would under the appropriate denaturation and annealing conditions, come together to form a double stranded nucleic acid molecule having at least one single stranded 5' or 3' terminal sequence.

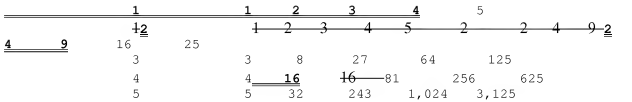
In one embodiment of the method, the nucleic acid components can be linked simultaneously to form the nucleic acid construct. Simultaneous assembly involves the incubation of nucleic acid components required for the assembly of a construct of interest, in the same reaction mixture. In another embodiment of the method, the nucleic acid components can be linked sequentially to form the nucleic acid construct. Sequential assembly is performed in a series of different reaction mixtures. This unique attribute lends itself to the automation of construct assembly. The method of the invention uses, preferably, attachment to a solid support as a starting point in the assembly of a series of nucleic acid components, in a defined order, to form a multicomponent nucleic acid construct. The method can be used to produce nucleic acid~~,~~ constructs which are functional as assembled (e.g. vectors) or constructs which are used as subcomponents for the assembly of functional constructs (e.g. genes or gene fragments attached to regulatory elements required for the expression of the gene or the gene fragment).

In still another embodiment, the method of the invention can be used to synthesize a group of nucleic acid constructs in which one or more of the components is substituted, in each of the constructs, with a different component, having the same functionality or characteristic utility. In this way the function of the different components can be evaluated and an optimal construct for a particular application identified. For example, as Table I shows, a cloning vector comprised of five different categories of nucleic acid components (e.g. origin of replication, resistance gene, promoter, etc.) might be designed so that users could choose amongst 5 different choices of nucleic acid components within each category. The number of permutations, or possible vector combinations, which are achievable from these ~~25~~**5** components is 3,125. Thus, it can be easily shown that a huge variety of different nucleic acid constructs which potentially address a wide range of highly specific user needs can be synthesized using a very small number of nucleic acid components.

TABLE I

Permutation of Constructs

Number of Different Nucleic Acid Component Categories	No. of Components within a Category 1	~~2~~**2**	~~2~~**3**	~~3~~ ~~4~~**4**	**5**



In preferred embodiments, the subject method is used to produce a combinatorial nucleic acid library having at least 4 different multicomponent constructs, more preferably at least 8, and even more preferably at least 16.

In another embodiment, the nucleic acid components may be covalently or non-covalently modified prior to or following assembly of the nucleic acid multicomponent construct. For instance, sites for the attachment of small biological molecules or macromolecular biological molecules, including proteins or carbohydrates may be added, enabling users to synthesize constructs having altered biological properties.

The method of this invention is particularly suitable for the construction of nucleic acid vectors. These include plasmid, viral, or phage vectors, or ~~yeast~~ artificial chromosomes**, e.g., such as for use in bacteria, yeast or mammalian cells**. The vector can be a cloning or expression vector and can be used for the expression of cDNA or genomic libraries, genes or gene fragments, mutagenized genes, recombined fusion genes, and artificial genes. The constructs can be employed in prokaryotic, eukaryotic (mammalian or non-mammalian) expression, construction of unique cDNA libraries, protein, **as wll as as** antibody and peptide **display libraries (e.g.** phage **peptide** display libraries **and mammalian peptide display libraries**). The constructs can further be employed in gene transfer, gene therapy, and the creation of transgenic organisms.

According to the method, the vector is assembled from nucleic acid components ~~encoding~~**providing** a single functionality or multiple functionalities~~. At a minimum,~~ **as appropriate, In one embodiment, a** nucleic acid components ~~encoding~~**including** an origin of replication, a selectable marker and an insert of interest**, such as an open reading frame,** are used. Depending on the type of vector desired, nucleic acid components encoding other vector functions may also be incorporated (e.g. a promoter, a transcription or translation regulatory element, etc.). An expression vector can be produced using a nucleic acid component encoding a structural gene or gene fragment of interest and additional nucleic acid components encoding regulatory elements required for expression of the gene. For example, a cDNA library expression vector is produced using nucleic acid components encoding a collection of cDNA molecules derived from poly(A)+ mRNA. Importantly, the optimization procedure of interchanging nucleic acid components described above can be used to create an optimal vector for a particular application.

B. General Methods Used in the Practice **of** the Invention

The practice of the present invention will employ, unless otherwise indicated, conventional techniques of recombinant DNA, molecular biology, cell biology, cell culture, transgenic biology, microbiology, and immunology, which are within the skill of the art. Such techniques

are described in the literature. See, for example, Molecular Cloning A Laboratory Manual, 2nd Ed., ed. by Sambrook, Fritsch and Maniatis (Cold Spring Harbor Laboratory Press: 1989).

i) Nucleic Acid Purification

Nucleic acid isolation procedures are performed essentially as described in Maniatis et al. Common nucleic acid isolation procedures involve cell lysis by detergents, protease treatment, and CsCl gradient purification. The latter step can be alternatively performed using commercially available binding matrices in the form of columns (e.g. Qiagen Kit).

ii) Oligonucleotide Synthesis

Oligonucleotide synthesis from the phosphoramidite versions of the nucleosides that DNA and RNA are composed from may be carried out on commercially available solid phase oligonucleotide synthesis machines (Needham-VanDevanter, D. R., et al., Nucleic Acids Res., 12:6159-6168, 1984), or chemically synthesized using the solid phase phosphoramidite triester method described by Beaucage et al., (Beaucage et al., Tetrahedron Letts. 22, No. 20:1859-1862, 1981).

Oligonucleotides are purified prior to use. Purification of oligonucleotides can be performed using reverse phase or anion-exchange HPLC and may also be carried out by denaturing or native ~~polyacrylamide~~**polyacrylamide** gel electrophoresis. Following purification, oligonucleotides can be phosphorylated using T4 polynucleotide kinase. As used herein, the term "T4 polynucleotide kinase" refers to the enzyme catalyzing the transfer of the terminal (~~γ~~**gamma**) phosphate of ATP to the 5' OH-terminus of a nucleic acid molecule.

iii) Restriction Enzyme Digestion

The procedures concerning the use of restriction enzymes, their nucleotide specificity and the appropriate reaction conditions are known to those skilled in the art and readily available. The amounts of enzyme and DNA, the buffer and ionic concentrations, and the temperature and duration of the reaction will vary depending upon the specific application as described in Maniatis et al.

iv) Ligation

Ligation of single stranded terminal sequences ~~is~~**can be** catalyzed by a ligase. As used herein, the term "ligase" refers to an enzyme that is capable of joining the 3' hydroxyl terminus of one nucleic acid molecule to a 5' phosphate terminus of a second nucleic acid molecule to form a single molecule. Most preferably, the T4 DNA ligase is used.

Ligation is carried out at 12.degree. C. to 16.degree. C. to maintain a balance between annealing of the terminal sequences and activity of the enzyme. An appropriate buffer containing the ATP cofactor required by the ligase, is used. When an enzymatic reaction, such as a ligation, is being conducted, it is preferable to provide the elements required for such a reaction in excess, such that the ability to achieve the desired ligation is not limited by the concentration of the elements.

y) PCR Amplification

The use of PCR is well known in the art and is described in U.S. Pat. No. 4,683,202, the contents of which are expressly incorporated herein by reference. The technique is described in several general sources, which provide adequate guidance to one of skill in the art, including Maniatis et al. and "PCR Protocols, A Guide to Methods and Applications" (Innis et al. eds.), Academic Press, San Diego, Calif. ~~1990.~~ **,1990. However, other amplification techniques, such as strand displacement amplification (SDA), are known and can be useful in the practice of the subject methods.**

C. Synthesis of the Nucleic Acid Component Terminal Sequences

Important elements of the method of the invention are terminal sequences, which are required for the efficient assembly of multiple nucleic acid components. The preferred type of terminal sequence is non-palindromic, even though palindromic terminal sequences or a mixture of palindromic and non-palindromic terminal sequences could be used. **That is, the single-stranded overhang is not a product of a restriction enzyme** The benefits of using non-palindromic terminal sequences are that there is no possibility of self-ligation and, in general, the terminal sequences may be designed so that only a single pair of terminal sequences are complementary and will exclusively anneal with each other. The size of the terminal sequences may be varied, but in general, the larger the size of the terminal sequence, the greater the fidelity of annealing specific and complementary terminal sequences within a mixture of numerous other terminal sequences. However, in certain preferred embodiments, the terminal sequences are about 6 to about 20 nucleotides in length, **more preferably** about 6 to about 15 nucleotides in length ~~or~~**, and even more preferably** about 6 to about 10 nucleotides in length.

Terminal **single-stranded overhang** sequences may be **provided at** either **or both of the** 5' or 3' ~~or both (see FIG. 2),~~ **ends of the nucleic acid component (e.g., see FIG. 2). Preferably, at least one of the terminal sequences is a non-palindromic overhang.** The primary constraint is that a 5' terminal sequence, in general, must anneal with a complementary 5' terminal sequence or an oligonucleotide (or series of oligonucleotides) which provide a complementary 5' terminal sequence. Likewise, a 3' terminal sequence must, in general, anneal with either a complementary 3' terminal sequence or an oligonucleotide (or series of oligonucleotides) which provide a complementary 3' terminal sequence. **The use of a bridging oligonucleotide which is complementary to both a 3' overhang on a first nucleic acid component and a 5' vector overhang on a second nucleic acid component is shown in FIG. 2. D. A second strategy for joining two nucleic acid components is by means of a "semi-cruciform" bridging element as shown in FIG. 3. In this embodiment of the joining method, the bridging element is comprised of two separate but partially complementary oligonucleotides which bridge the 5' and 3' overhangs of adjacent nucleic acid components. The semi-cruciform bridged nucleic acid components can then be covalently joined by a ligation step as shown.**

Terminal sequences may be synthesized by using a number of different methods including, without limitation, the following:

(1) Adaptors may be ligated to restriction enzyme digested nucleic acid components. These adaptor molecules are composed of synthetic oligonucleotides which are designed to be complementary at one end with a restriction enzyme digested nucleic acid molecule and the other end containing a single stranded terminal sequence, preferably non-palindromic.

(2) Oligonucleotide primers, which contain one or more synthetic uracil residues, may be utilized to PCR-amplify a fragment, followed by uracil DNA glycosylase treatment, resulting in 3' terminal sequences, a method described in U.S. Pat. No. 5,137,814, the contents of which are expressly incorporated herein by reference. "Uracil DNA glycosylase" (UDG), a term of art, refers to an activity which cleaves the glycosidic bond between the base uracil and the sugar deoxyribose, only when the monomeric nucleotide dUTP is incorporated into a DNA molecule, resulting in incorporation of a deoxyuridine moiety (Duncan, B. in The Enzymes 14:565 (1981, ed.: Boyer P.). An enzyme possessing this activity does not act upon free dUTP, free ~~deoxyuridine~~**deoxyuridine**, or RNA (Duncan, supra). The action of UDG results in the production of an "abasic" site. The enzyme does not, however, cleave the phosphodiester backbone of the nucleic acid component. Most preferably, the phosphodiester backbone at an abasic site may be cleaved through the use of an endonuclease specific for such substrates. A preferred enzyme for this purpose is the E. coli enzyme, Endonuclease IV. Most preferably, Endonuclease IV is used in conjunction with UDG to remove dU residues from a nucleic acid component.

(3) 5' terminal sequences may be generated in PCR products by using PCR oligonucleotide primers containing alkane diol derivatives, a method described in U.S. Pat. No. 5,426,039, the contents of which are expressly incorporated herein by reference. These same type of modified primers may be used when using non-PCR amplification methods, resulting in the same type of unique terminal sequences as defined by these primers.

(4) Suitable nucleic acid component 3' terminal sequences may be generated by PCR with phosphorothioate substituted oligonucleotides followed by exonuclease digestion. A particularly preferred method for the synthesis of the nucleic acid terminal sequence component is by a specialized pcr amplification and cloning technique as shown in FIG. 4. The technique produces single-stranded 3' DNA overhangs suitable for joining the nucleic acid components of the present invention. In a preferred embodiment, the 3' overhangs are 12 to 15 nucleotides long. The technique involves incorporating phosphorothioate-nucleotide substituted oligonucleotide primers onto the ends of a target nucleic acid component by means of pcr amplification. The oligonucleotide primers are designed to incorporate the phosphorothioate nucleotide linkage at a specific site within the primers. The substituted linkage is resistant to exonuclease cleavage and so exposure of the nucleic acid component amplification product to an appropriate 5' exonuclease establishes a discrete and stable proximal end to the 3' single stranded overhang on the undigested strand. Suitable exonuclease enzymes include lambda exonuclease, which possesses a 5' to 3' exonuclease activity but no 3' to 5' exonuclease activity. The resulting exonuclease digested phosphorothioate substituted oligonucleotide amplification product will contain unique 3' single-stranded overhangs corresponding in sequence to the complement of the sequence 5' proximal to the phosphorothioate linkage of the primer. Complementary 3' overhangs can be incorporated into a cloning vector which would allow for specific

annealing with 3' overhangs of the fragment to be cloned.

In one embodiment of the invention, this 5' phosphorothioate proximal portion of the primer corresponds to a sequence in the target nucleic acid component. In a preferred embodiment, however, this 5' phosphorothioate proximal portion of the primer is uniquely introduced into the pcr product by the amplification process itself. The incorporation of standard restriction endonuclease recognition elements onto the 5' ends of pcr amplification primers is a well established method of facilitating the cloning of pcr-generated sequences. The standard method involves cleavage of the pcr product with the corresponding restriction endonuclease to create a 5' or 3' overhang. The phosphorothioate-substituted primer/exonuclease digestion method is particularly advantageous over the standard restriction site method in that virtually any sequence (palindromic or nonpalindromic) can be introduced into the 3' overhang. Thus the 3' overhang can be specifically tailored to the application.

Generally, in joining two nucleic acid components in a particular orientation, the pcr primers are designed so that the ends to be joined correspond to complementary 5' regions of the phosphorothioate linkage substituted primers (see FIG. 3). These regions of 5' complementarity are converted into complementary 3' overhangs following per amplification and exonuclease treatment as described above. These 3' overhangs can be uniquely adapted so that the various nucleic acid components are joined in only one possible polarized configuration.

The flexibility afforded by this unique manner of introducing 3' overhangs onto any nucleic acid component is particularly suited to the joining of those nucleic acid components which directly affect the expression or form of the insert of interest. This is because expression signal sequences are often affected by the sequences immediately surrounding them. Furthermore, the proteins expressed by nucleic acid components encoding fusion polypeptide domains are often influenced by the polypeptides they are fused to due to the lack of predictability of protein folding and steric relationships between the two joined polypeptides. The immediate invention provides a means of introducing virtually any sequence in these positions and thereby allows the maintenance of optimized function of individualized elements even as they are shuffled in numerous combinations with other elements. This flexibility is particularly suited to the joining of polypeptide-encoding nucleic acid components because the unique 3' overhangs can be uniquely designed to allow for: maintenance of the appropriate reading frame across a polypeptide fusion junction; introduction of small polypeptide-encoding domains, such as the FLAG epitope, between polypeptide-encoding nucleic acid components; and the introduction of particular "floppy" polypeptide-sequences, such as poly-glycine, between polypeptide elements such that the resulting joined polypeptides fold independently and retain independent functions.

In one embodiment, the resulting nucleic acid components containing the terminal sequences, can be isolated by agarose or acrylamide gel electrophoresis followed by elution of the nucleic acid components from the agarose or acrylamide matrix. The two most common ways of elution are either soaking in an appropriate buffer or electroelution, both described in Maniatis et al. Both methods are effective, but soaking is often the method of choice because it is inexpensive,

easy and can be accomplished without monitoring. Kits for the purification of nucleic acids from gel matrices may also be used (e.g. "Compass Kit", American Bioanalytical). In another embodiment, the resulting nucleic acid components, containing the terminal sequences, can be purified using reverse phase or anion-exchange HPLC.

D. Covalent Assembly of the Nucleic Acid Components

In the method of the invention, the various nucleic acid components are designed so that each component ~~contains~~**may contain** specific and unique terminal sequences at either end. Each terminal sequence ~~is~~**can be** designed to anneal and base pair with a unique complementary terminal sequence residing on a separate nucleic acid component. A series of specific annealing reactions occur between complementary terminal sequences. This results in the assembly of a larger nucleic acid multicomponent construct having a defined relative order and orientation for all the components.

According to the method of the invention, the various nucleic acid components can be linked via, without limitation, the following:

(1) Annealing of 5' complementary terminal sequences in two separate nucleic acid components (FIG. 2B).

(2) Annealing of 3' complementary terminal sequences in two separate nucleic acid components (FIG. 2C).

(3) Annealing of an oligonucleotide bridge with complementary 5' and 3' terminal sequences in two separate nucleic acid components (FIG. 2D).

(4) Annealing of an adaptor molecule with complementary 5' or 3' terminal sequences in two separate nucleic acid components (FIG. 2E).

(5) Annealing of a semi-cruciform bridge comprising two partially complementary oligonucleotides with complementary 5' and 3' terminal sequences in two separate nucleic acid components (FIG. 3).

The fidelity of assembly of the nucleic acid multicomponent construct depends upon a number of factors, including, without limitation, the following: ~~(~~1) The number of different nucleic acid components, ~~(~~2) The size of the terminal sequences, ~~(~~3) The way anealing occurs, ~~(~~4) The annealing conditions, ~~(~~5) The nucleotide sequence within the terminal sequences.

In certain embodiments, the generation of multicomponent constructs can be monitored by incorporation of nucleic acid components which include portions of a genetic element(s) which, for its functionality to be apparent, requires the presence of two or more nucleic acid components in the final construct. For example, a marker gene can be split between two nucleic acid components, and is only detected when the gene is recapitulated by successful ligation of the two nucleic acid components.

In a preferred embodiment of the invention, three or more nucleic acid components are used for the production of a nucleic acid construct. Preferably three, four, five, or six nucleic acid components are used and more preferably three to eight nucleic acid components are used. Using the method of the invention, the various nucleic acid components can be incubated either simultaneously or in a step-wise fashion, to form nucleic acid multicomponent constructs which can be either functional as assembled or which can be used as subcomponents for the assembly of functional constructs. Three or more nucleic acid components may be linked to form a nucleic acid multicomponent construct. Functional constructs may be assembled from such nucleic acid multicomponent constructs, with each multicomponent construct essentially performing as a single nucleic acid component in the assembly of a functional construct. Nucleic acid multicomponent constructs would be preferably employed when there are a large number of different nucleic acid components requiring assembly, when there are non-unique terminal sequences within a group of different nucleic acid components, or when the size of the final assembled functional construct is very large. Nucleic acid multicomponent constructs may also be used in repetitive cloning experiments or in the design of assembly reactions which are repetitive or otherwise simplified.

Typically, the nucleic acid components would include an appropriately phosphorylated terminal sequence, suitable for ligation to a separate nucleic acid component. The nucleic acid components are incubated under appropriate conditions that allow for efficient annealing of the complementary terminal sequences. Appropriate annealing conditions are described in Maniatis et al. In a particularly preferred embodiment of the invention, the nucleic acid components are incubated in equimolar concentrations, heated to 65.degree. C., and then cooled down slowly to 25.degree. C. Temperatures ranging from ~~60.degree.~~**60** to 75.degree. C. may be used depending on the size of the terminal sequences employed.

In certain ~~preferred~~ embodiments of the invention, the nucleic acid components are treated with a ligase enzyme to ligate the nucleic acid components and produce a nucleic acid construct. Preferably a T4 DNA ligase is used, even though the E. coli ligase may also be used for certain applications. In another embodiment of the method of the invention, ligation of the different nucleic acid components may not be necessary prior to transferring the assembled nucleic acid construct into the appropriate biological or experimental system.

In yet another embodiment, the combinatorial method can be carried out in a manner that utilizes flanking intronic sequences to generate a combinatorial library of multicomponent constructs. As illustrated schematically in FIG. 5, the combinatorial event takes place at the DNA level through annealing of complementary sequences within intronic portions of the nucleic acid components. Briefly, double-stranded nucleic acid components are generated which include an "exonic sequence" and flanking intron fragments. That is, intronic sequences flanking the 5' end of an exon module represent a 3' fragment of an intron. As described herein, 5' and 3' non-palindromic overhangs are generated in the nucleic acid components. Annealing of the non-palindromic terminal sequences, therefore, mediates concatenation of the component to one and other through basepairing. In the exemplary illustration of FIG. 5, the exon sequences are flanked by domains IV-VI of an autocatalytic group II intron at one end, and domains I-IV at the other. A library of combinatorial units representative of a number of different exons is generated. Upon annealing of the non-

palindromic terminal sequences, the sequences corresponding to domain IV at the 3' end of one unit anneal with the complementary domain IV sequences at the 3' end of another unit, resulting in concatenation of combinatorial units (see FIG. 5).

The resulting combinatorial genes can be subsequently cloned into an expression vector. In one instance, 5' terminal and 3' terminal combinatorial units can be used and the double-stranded genes can be amplified using PCR anchors which correspond to sequences in each of the two terminal units. The PCR primers can further be used to add restriction endonuclease cleavage sites which allow the amplified products to be conveniently ligated into the backbone of an expression vector. Upon transcription of the combinatorial gene, the intronic RNA sequences will drive ligation of the exonic sequences to produce an intron-less transcript. In this manner, the subject GEOS method can be used to introduct non-palindromic overhangs in the nucleic acid components without altering the coding sequence.

While FIG. 5 demonstrates one embodiment which utilizes group II introns, the combinatorial process can be carried out in similar fashion using either group I intron sequences or nuclear pre-mRNA intron sequences.

As used herein, the terms "exon" and "exonic sequence" denotes nucleic acid sequences, or exon "modules", that can, for instance, encode portions of proteins or polypeptide chains, such as corresponding to naturally occurring exon sequences or naturally occurring exon sequences which have been mutated (e.g. point mutations, truncations, fusions), as well as nucleic acid sequences from "synthetic exons" including sequences of purely random construction. However, the term "exon", as used in the present invention, is not limited to protein-encoding sequences, and may comprises nucleic acid sequences of other function, including nucleic acids of "intronic origin" which give rise to, for example, ribozymes or other nucleic acid structure having some defined chemical function.

E. Preparation of Synthetically or Covalently Modified Nucleic Acid Components

A unique feature of the method of the invention is that, since nucleic acid components may be made synthetically, any nucleic acid component may be altered or modified to contain one or more modifications (i.e., handles). Handles may act as sites of attachment for small biological molecules or macromolecular biological molecules, including proteins or carbohydrates. They may also serve as sites of attachment for chemically synthesized, non-biological molecules. The method of the invention, therefore, enables users to synthesize constructs having altered biological properties.

Modifications which could be performed on nucleic acid components include, without limitation, the following: Modification of nucleic acid residues, biotinylation, fluorescent tagging, incorporation of polypeptide nucleic acids (PNA), covalent or non-covalent conjugation of proteins involved in nucleic acid modification **(e.g. through the use of activated boronic acid moieties)**, including enzymes, covalent or non-covalent conjugation of proteins or other components or ions which enable the recognition and binding of specific molecular targets, including haptens.

Modification of nucleic acid residues can be performed by a variety of art known techniques. The simplest method for performing oligonucleotide directed mutagenesis is by enzymatic primer extension (PCR). In this method, an oligonucleotide primer is designed that carries the mutation of interest flanked by 10 to 15 nucleotides of wild-type sequence. This "mutagenic" oligonucleotide can then be used in a PCR reaction along with an oligonucleotide primer containing one or more synthetic uracil residues or alkane diol derivatives to create the nucleic component of interest. The types of mutations that can be made by this approach range from single nucleotide substitutions to deletions or insertions, limited only by the size of the oligonucleotide primer needed.

The synthesis of biotinylated nucleotides is well known in the art and was first described by Langer et al. (PNAS 78:6633-37, 1981). Biotin, a water soluble vitamin, is covalently attached to the C5 position of the pyrimidine ring via an allylamine linker arm. Biotinylation of DNA can be achieved by either nick translation, adapted successfully to incorporate biotinylated nucleotides (biotin-11 and biotin-16-dUTP, biotin-14-dATP), or random-priming using biotinylated octamers. Biotinylated nucleic acid molecules can be prepared from biotin-NHS (N-hydroxy-succinimide) using techniques well known in the art (e.g., biotinylation kit, Pierce Chemicals, Rockford, Ill.).

Fluorescent tagging of nucleic acid molecules can be performed using techniques well known in the art (e.g. using the Fluore-dUTP Labelling Mix by Pharmacia) Examples of suitable fluorescent materials include umbelliferone, fluorescein, fluorescein isothiocyanate, rhodamine, dichlorotriazinylamine fluorescein, dansyl chloride or phycoerythrin.

In an embodiment of the invention, synthetic oligonucleotides are used that contain polypeptide nucleic acids or functional groups like primary amines, sulfhydryls, disulfides, and any other group typically used for conjugation of haptens, proteins, enzymes or antibodies.

~~Assembly of Vector Constructs~~

F. GEOS Nucleic Acid Components

Another aspect of the invention pertains to the assembly of vectors, preferably expression vectors, using a series of interchangeable nucleic acid components **or "vector elements"**. As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to which it has been linked. Certain vectors are capable of autonomous replication in a host cell into which they are introduced (e.g., bacterial vectors having a bacterial origin of replication and episomal mammalian vectors). Other vectors (e.g., non-episomal mammalian vectors) are integrated into the genome of a host cell upon introduction into the host cell, and thereby are replicated along with the host genome.

The invention provides methods for the rapid assembly of multi-element vector systems in which one or more of the various elements of the expression system can be varied so as to create a collection of candidate expression clones from which optimization of expression can be achieved by appropriate screening or selection techniques. Those elements to be

included in any vector system will depend upon the system's intended use. Furthermore those elements of the multi-element system to be systematically varied will further depend upon both its intended use as well as the nature of difficulties anticipated in achieving suitable expression in a given context. The method is particularly suited to the production of complex pools of expression vectors in which one or more nucleic acid components are varied. Such pools may serve as "libraries" of candidate expression vectors from which particular species may be selected on the basis of preferred properties. Alternatively, the heterogeneous pools may be directly useful in applications for which expression of heterogeneous populations of related species may be desirable (e.g. DNA vaccines).

The vector elements can be roughly divided into two categories: those which directly affect the expression or form of the insert of interest and those which do not (i.e. vector "backbone" elements). Examples of the former include: promoter elements, enhancer elements, transcription initiation elements, transcription termination elements, polyadenylation signal elements, intronic elements, translation initiation elements, epitope tag elements, and various polypeptide fusion elements. Examples of the latter typically include: selectable marker elements, origin of replication elements, integration elements, integration-promoting factor elements, and chromosomal structural elements such as centromeric attachment elements and telomeric elements.

Preferred promoter vector elements of the present invention include both eukaryotic and prokaryotic promoter elements. Preferred prokaryotic promoter elements of the invention include those which carry optimal -35 and -10 (Pribnow box) sequences for transcription by RNA polymerase in Escherichia coli. Some prokaryotic promoter elements will contain overlapping binding sites for regulatory repressors (e.g. the Lac, and TAC promoters, which contain overlapping binding sites for lac repressor thereby conferring inducibility by the substrate homolog IPTG). Examples of prokaryotic genes from which suitable promoter sequences may be obtained include E. coli lac, ara, and trp. Prokaryotic viral promoter elements of the present invention include lambda phage promoters (e.g. P.sub.RM and P.sub.R), T7 phage promoter elements, and SP6 promoter elements. Eukaryotic promoter vector elements of the invention include both yeast (e.g. GAL1, GAL10, CYC1) and mammalian (e.g. promoters of globin genes and interferon genes). Preferred eukaryotic promoter vector elements include viral gene promoters such as those of the SV40 promoter, the CMV promoter, herpes simplex thymidine kinase promoter, as well as any of various retroviral LTR promoter elements (e.g. the MMTV LTR). Still more preferred eukaryotic promoter vector elements contain both a control region in addition to a promoter so that the resulting construct can be derepressed by a suitable inducing agent. Examples of such control regions include those which bind the tetracycline resistance repressor, specific forms of which can cause regulated activation or repression of a linked promoter in response to tetracycline.

Promoter vector elements of the invention include both inducible and repressible promoters. The inducible promoters of the present invention include those which are capable of functioning in a eukaryotic host organism. Preferred embodiments include naturally occurring yeast and mammalian inducible promoters as well as synthetic promoters designed to function in a eukaryotic host as described below. The important

functional characteristic of the inducible promoters of the present invention is their ultimate inducibility by exposure to an environmental inducing agent. Appropriate environmental inducing agents include exposure to heat, various steroidal compounds, divalent cations (including Cu,sup.+2 and Zn,sup.+2), galactose, tetracycline, IPTG (isopropyl b-D thiogalactoside), as well as other naturally occurring and synthetic inducing agents and gratuitous inducers. In certain modes of the invention, the environmental inducing signal can correspond to the removal of any of the above listed agents which are otherwise continuously supplied in the uninduced state (see the tTA based system described below for example). The inducibility of a eukaryotic promoter can be achieved by either of two mechanisms included in the method of the present invention. Suitable inducible promoters can be dependent upon transcriptional activators which, in turn, are reliant upon an environmental inducing agent. Alternatively the inducible promoters can be repressed by a transcriptional repressor which itself is rendered inactive by an environmental inducing agent. Thus the inducible promoter can be either one which is induced by an environmental agent which positively activates a transcriptional activator, or one which is derepressed by an environmental agent which negatively regulates a transcriptional repressor. We note here that the latter class of inducible promoter systems defines transcriptional repressors and corresponding negative cis regulatory elements which can also find use as the repressors and corresponding repressible promoters of the present invention as described in section 4.3.2.

The inducible promoters of the present invention include those controlled by the action of latent transcriptional activators which are subject to induction by the action of environmental inducing agents. Preferred examples include the copper inducible promoters of the yeast genes CUP1, CRS5, and SOD1 which are subject to copper-dependent activation by the yeast ACE1 transcriptional activator (see e.g. Strain and Culotta (1996) Mol Gen Genet 251: 139-45; Hottiger et al. (1994) Yeast 10: 283-96; Lapinskas et al. (1993) Curr Genet 24: 388-93; and Gralla et al. (1991) Proc. Natl. Acad. Sci. USA 88: 8558-62). Alternatively, the copper inducible promoter of the yeast gene CTT1 (encoding cytosolic catalase T), which operates independently of the ACE1 transcriptional activator (Lapinskas et al. (1993) Curr Genet 24: 388-93), can be utilized. The copper concentrations required for effective induction of these genes are suitably low so as to be tolerated by most cell systems, including yeast and Drosophila cells. Alternatively, other naturally occurring inducible promoters can be used in the present invention including; steroid inducible gene promoters (see e.g. Oligino et al. (1998) Gene Ther. 5: 491-6); galactose inducible promoters from yeast (see e.g. Johnston (1987) Microbiol Rev 51: 458-76; Ruzzi et al. (1987) Mol Cell Biol 7: 991-7); and various heat shock gene promoters. Many eukaryotic transcriptional activators have been shown to function in a broad range of eukaryotic host cells, and so, for example, many of the inducible promoters identified in yeast can be adapted for use in a mammalian host cell as well. For example, a unique synthetic transcriptional induction system for mammalian cells has been developed based upon a GAL4-estrogen receptor fusion protein which induces mammalian promoters containing GAL4 binding sites (Braselmann et al. (1993) Proc Natl Acad Sci USA 90: 1657-61). These and other inducible promoters responsive to transcriptional activators which are dependent upon specific inducing agents are suitable for use with the present invention.

The inducible promoters of the present invention also include those which are repressed by repressors which are subject to inactivation by the action of environmental inducing agents. Examples include prokaryotic repressors which can transcriptionally repress eukaryotic promoters which have been engineered to incorporate appropriate repressor-binding operator sequences. Preferred repressors for use in the present invention are sensitive to inactivation by physiologically benign inducing agent. Thus, where the lac repressor protein is used to control the expression of a eukaryotic promoter which has been engineered to contain a lacO operator sequence, treatment of the host cell with IPTG will cause the dissociation of the lac repressor from the engineered promoter and allow transcription to occur. Similarly, where the tet repressor is used to control the expression of a eukaryotic promoter which has been engineered to contain a tetO operator sequence, treatment of the host cell with IPTG will cause the dissociation of the tet repressor from the engineered promoter and allow transcription to occur.

In a preferred embodiment of the invention, the repressor of the inducible promoter is synthesized as a ubiquitin fusion protein conforming to the formula ubiquitin-X-repressor. This can be achieved using the ubiquitin fusion vector systems designed to confer inducible proteolytic sensitivity to the target gene polypeptide as described below. Thus it will be appreciated by the skilled artisan that a rapid induction of a repressible promoter can be achieved by simultaneously delivering an environmental inducing agent which causes dissociation of the repressor from the repressed inducible promoter, and simultaneously promoting the destruction of that repressor by N-end rule directed proteolysis. Degradation of the repressor prevents rebinding to the operator which can result in decreased inducibility of the repressible promoter--a problem which has been recognized in the art (see Gossen et al. (1993) TIBS 18: 471-5). Furthermore, this aspect of the invention can be utilized independently of the targeted shut-off of a gene, to generally increase the inducibility of a eukaryotic expression system which is subject to repression by a repressor. Thus the present invention further provides improved methods for inducible expression of endogenous or heterologous genes in a eukaryotic cell.

As suggested above, the inducible promoters of the present invention include those which are not naturally occurring promoters but rather synthetically derived inducible promoter systems which may make use of prokaryotic transcriptional repressor proteins. The advantage of using prokaryotic repressor proteins in the invention is their specificity to a corresponding bacterial operator binding site, which can be incorporated into the synthetic inducible promoter system. These prokaryotic repressor proteins have no natural eukaryotic gene targets and affect only the effector of suppression gene which is put under the transcriptional control of the inducible synthetic promoter. This system thereby avoids undesirable side-effects resulting from unintentional alteration of the expression of nontargeted eucaryotic genes when the inducible promoter is induced. A preferred example of this type of inducible promoter system is the tetracycline-regulated inducible promoter system. Various useful versions of this promoter system have been described (see Shockett and Schatz (1996) Proc. Natl. Acad. Sci. USA 93: 5173-76 for review). As suggested above, these tetracycline-regulated systems generally make use of a strong eucaryotic promoter, such as human cytomegalovirus (CMV) immediate early (IE) promoter/enhancer and a tet resistance operator (tetO) which is bound by the tet repressor protein. In a preferred

embodiment, the system involves a modified version of the tet repressor protein called a reverse transactivator (rtTA, or rtTA-nls, which contains a nuclear localization signal) which binds tetO sequences only in the presence of the tet derivatives doxycycline or anhydrotetracycline. Using this system, a synthetic human CMV/IE-tetO-promoter driven construct could be induced by 3 orders of magnitude in 20 hrs by the addition of the tet derivatives (see Gossen et al. (1995) Science 268: 1766-9). Thus this system can be used to make the effector of suppression genes of the present invention inducible in response to the delivery of tetracycline derivatives to the targeted eucaryotic cell. Alternatively, a tet repressor fused to a transcriptional activation domain of VP16 (tTA) can be used to drive expression of the inducible promoter of the present invention. In this instance, transcriptional activation of a synthetic human CMV/IE- tetO-promoter driven construct is achieved by the removal of tetracycline since the tTA activator only binds to the tetO in the absence of tet (see Gossen and Bujard (1992) Proc. Natl. Acad. Sci. USA 89: 5547-51). Other synthetic inducible promoter systems are also available for use in the present invention. For example, a lac repressor-VP16 fusion which exhibits a "reverse" DNA binding phenotype (i.e., analogous to rtTA described above, it only binds the lacO operator sequence in the presence of the inducer IPTG) (see Lambowitz and Belfort (1993) Annu Rev Biochem 62: 587-622). This particular synthetic inducible promoter is approximately 1000-fold inducible in the presence of IPTG. Since neither the tet repressor gene nor the lac repressor gene occurs naturally in a eukaryotic cell, systems involving synthetic inducible promoter constructs such as these rely on the further delivery of an expressible copy of the appropriate prokaryotic repressor gene. Suitable expression cassettes for this purpose are readily available for heterologous expression in many different eukaryotic cells including various yeast species and mammalian cells.

Another vector element category for use in the present invention is a eukaryotic enhancer element. Preferred eukaryotic enhancer elements include those which are tissue and/or developmentally specific. Incorporation of such tissue-specific vector elements by GEOS methodology can be particularly useful in gene therapy applications in which only a specific human tissue is targeted for expression. Other enhancers elements for use in the present invention include viral transcriptional enhancers such as the SV40 enhancer, the polyoma virus enhancer, and retroviral LTR enhancers. Preferred viral enhancer elements exhibit strong cell-type preference in transcriptional activity. Examples include enhancer elements within the Moloney MuSV, which are regulated by the glucocorticoid dexamethasone and a hepatitis B enhancer associated with a hepatitis surface antigen coding sequence which is specifically active in human liver cells.

Yet another vector element of the present invention is fusion polypeptide-encoding element which can be fused to the coding region of any insert gene of interest. For example, in certain applications it is useful to be able to mark a particular gene product with a tag so that the localization and function of the gene product can be easily monitored. A particularly preferred version of a molecular tag fusion polypeptide element is the green fluorescent protein or GFP (see e.g. Misteli and Spector (1997) Nature Biotechnology 15: 961-4; and Gerdes and Kaether (1996) FEBS Lett 389: 44-7 for review). This fusion tag polypeptide emits green (approximately 510 nm wavelength) light upon excitation by a particular wavelength of incident light (approximately 400 to 480 nm, depending upon the

form of GFP). Various versions of GFP coding sequences, including those whose codon usage has been humanized and those whose emission spectra have been "red-shifted," are commercially available and can be readily adapted to GEOS methodology. Applications of GFP include in situ localization of a linked gene of interest, as well as facile monitoring of expression and tropism in various cell mediated expression studies.

Still other vector element fusion polypeptides include beta-galactosidase and thioredoxin, as well as various affinity tags (e.g. polyHIS, which binds with high affinity to a columns to which Ni,sup.+2 is immobilized) and epitope tags, which are particularly suited to subsequent detection with corresponding antibodies (e.g. myc, FLAG, enterokinase, or hemagluttinin tags). Various epitope tag encoding sequences are available and can be readily adapted to GEOS methodology.

Yet another class of vector elements for use in the invention are transcriptional terminator elements which promote the formation of 3' ends in a mature RNA transcript. Such elements include prokaryotic terminator elements as well as eukaryotic terminator elements which function primarily as polyadenylation signals. Examples of the latter included various viral terminator elements such as the SV40 polyadenylation signal element and the beta-globin polyadenylation signal element.

In a particularly preferred embodiment of the invention, the vector elements comprise exonic and intronic units which can used to facilitate rearrangement of discrete polypeptide-encoding exon units. The GEOS methodology thereby facilitates the formation of peptide domain "shuffled" libraries encoded by exonic units which are linked by intronic units. Thus GEOS can be readily applied to the creation of unique proteins peptide domain shuffling and in vitro selection techniques. In one particular embodiment of the peptide domain shuffling application a pool of diverse peptide domain-encoding nucleic acid components is added to a GEOS vector assemblage. A compatible intronic unit is then added prior to the addition of another pool of diverse peptide domain-encoding exonic nucleic acid components. The resulting GEOS assembly increases in heterogeneity with successive addition of peptide domain-encoding components and provides a convenient source of diversity for subsequent screening or selection processes. GEOS methodology thereby enables the user to create synthetically "evolved" proteins comprised of discrete polypeptide domains which have been randomized and infinitely "shuffled."

A discussion of general considerations to be made in designing suitable vector systems is provided below.

One type of vector produced by the method of the invention is a minimal vector (referred to usually as a plasmid vector), which is basically a circular double stranded DNA loop into which additional DNA segments can be ligated. Another type of vector, produced by the method of the invention, is a vector capable of directing the expression of genes to which it is operatively linked. Such a vector is referred to herein as an "expression vector". The invention is intended to include the production of various forms of expression vectors, such as vectors derived from bacteriophage, including all DNA and RNA phage (e.g. cosmids), or viral vectors derived from: (a) all eukaryotic viruses, such as baculoviruses and retroviruses, (b) adenoviruses and adeno-

associated viruses, Herpes viruses, Vaccinia viruses and all single-stranded, double stranded and partially double stranded DNA viruses, (c) all positive and negative stranded RNA viruses, and (d) replication defective retroviruses.

Another type of vector produced by the method of the invention is a yeast artificial chromosome (YAC), which contains both a centromere and two telomeres, allowing YACs to replicate as small linear chromosomes. YACs can carry several hundred thousand base pairs of DNA, making them appropriate for genome mapping procedures. **Other artificial chromosomes include the PAC (P1 artificial chromosomes), BAC (bacterial artificial chromosomes) and MAC (mammalian artificial chromosomes), or exogenous extra-chromosomal components, e.g., derived from viruses and other cellular parasites.**

Each nucleic acid component involved in the assembly of a vector construct is intended to encode a specific biological functionality or multiple functionalities. For example, plasmid vectors generally contain several genetic elements such as the following: (a) an origin of replication, (b) a selectable marker element, (c) an insert of interest, for the insertion of genetic elements, such as a specific gene coding for a protein of interest.

The method of the present invention enables nucleic acid components to be synthesized to contain specific and unique terminal sequences such that annealing of complementary terminal sequences between different components will result in the generation of definable and specifically oriented constructs. A vector may be constructed by combining a set of nucleic acid components which provide all the necessary genetic elements required to generate a functional vector, while the unique terminal sequences on each component will determine the order in which all of the nucleic acid components are assembled relative to each other.

According to the method of the invention, individual nucleic acid components may be substituted with other components containing the same unique terminal sequences (see FIG. 1). For example, the plasmid origin of replication (ori) is a genetic element of a particular category, whose function is to initiate and regulate plasmid replication in bacteria, provide host range specificity, and regulate plasmid copy number and plasmid compatibility. This general functionality may be provided by a variety of different nucleic acid components within the ori category, including ori segments, ori genes or ori genetic elements. This invention allows for the synthesis and utilization of a series of different ori nucleic acid components, each having the same unique terminal sequences, which would enable users to rapidly and easily choose from a catalog of interchangeable ori nucleic acid components when designing and specifying a plasmid construct. Examples of origins of replication include the pMB1, p15A, 2 .mu., ColE1, psc101, F, R6K, R1, RK2, and .lambda.dv origins of replication.

"Selectable marker" as used herein, refers to the marker and to the nucleic acid encoding said marker. Selectable markers contemplated by the present invention include resistance to antibiotics such as ampicillin, tetracycline, chloramphenicol, kanamycin, neomycin, rifampicin, carnebicillin, streptomycin, and the like. The selectable markers also encompass resistance to drugs such as hygromycin and methotrexate, heavy metals such as cadmium, phage infection, and sensitivity to enzymes which affect calorimetric changes such as .beta.-galactosidase.

A vector may be assembled from multiple individual nucleic acid components, including, without limitation, nucleic acid components which incorporate one or more of the following: (a) origin of replication (bacterial, viral, phage, yeast, mammalian, eukaryotic), (b) selectable markers (antibiotic resistance, drug resistance, mutagenic resistance), (c) promoters (phage, bacterial, yeast, eukaryotic, mammalian), (d) regulatory elements or genes (repressors, enhances), (e) structural genes, (f) fragments of structural genes, (g) translational elements (Shine-Delgarno element, Kozak sequence), (h) terminators of transcription, (i) regulators of mRNA stability (degradation signals, translational regulators), (j) protein encoded elements specifying cellular location (leader sequence, KDEL, CAAX box, nuclear targeting elements), (k) recombination elements (Lox-CRE, M13 ori), (l) mutagenized genes, (m) protein domain encoded regions, (n) synthetic multiple cloning sites, (o) unique restriction enzyme or DNA cleavage sites, (p) site for covalent or non covalent attachment of a biological or chemical molecule (see "Handle").

In a preferred embodiment of the invention, an expression vector is produced. The expression vector produced by the method of the invention comprises nucleic acid components encoding one or more regulatory sequences, selected on the basis of the host cells to be used for expression, as well as the nucleic acid sequence to be expressed. The term "regulatory sequence" is intended to include promoters, enhancers and other expression control elements (e.g., polyadenylation signals). Such regulatory sequences are described, for example, in Goeddel; Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, Calif. (1990). Regulatory sequences include those which direct constitutive expression of a nucleotide sequence in many types of host cell and those which direct expression of the nucleotide sequence only in certain host cells (e.g., tissue-specific regulatory sequences). It will be appreciated by those skilled in the art that the design of the expression vector can depend on such factors as the choice of the host cell to be transformed, the level of expression of protein desired, etc. The expression vectors produced by the method of the invention can be introduced into host cells to thereby produce proteins or peptides, including fusion proteins or peptides.

The expression vectors produced by the method of the invention can be, for example, designed for expression of a gene of interest in prokaryotic or eukaryotic cells. For example, the expression vectors can be used for expression in bacterial cells such as E. coli, insect cells (using baculovirus expression vectors) yeast cells or mammalian cells. Suitable host cells are discussed further in Goeddel, Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, Calif. (1990). Alternatively, the expression vectors produced by the method of the invention can be transcribed and translated in vitro, for example using the T7 promoter regulatory sequences and the T7 polymerase. The expression vectors produced by the method of the invention can also be used to produce nonhuman transgenic animals. Furthermore, the nucleic acid vectors produced by the method of the invention can be used as gene therapy vectors. Gene therapy vectors can be delivered to a subject by, for example, intravenous injection, local administration (see U.S. Pat. No. 5,328,470) or by stereotactic injection (see e.g., Chen et al. (1994) PNAS 91:3054-3057). Vector constructs assembled using the method of the invention may also be used as templates to synthesize RNA using standard methods. Examples of RNA molecules which could be made, would include, without limitation, the following: mRNA, tRNA, rRNA, snRNA, hnRNA, viral or phage RNA, or modified RNA genes or genetic elements.

G. ~~A.~~ Assembly of Genomic and cDNA Libraries

i) Genomic Libraries

One aspect of the present invention pertains to the assembly of genomic libraries from individual nucleic acid components. Using the method of the invention, eukaryotic organism (e.g. viral) or prokaryotic organism (e.g. phage) genomes may be assembled in unique ways. The genome of an organism may be endonucleolytically or exonucleolytically cleaved using suitable restriction enzymes, followed by ligation of specific adaptor molecules, as described above.

For example, in one embodiment, the Lambda phage genome which is an approximately 50 kb double stranded DNA molecule encoding multiple genetic regulatory elements as well as approximately 30-40 structural genes, can be provided in the form of nucleic acid components. Toward this end, each of the Lambda phage genes, or groups of genes can be synthesized to contain unique terminal sequences so that these genes, or groups of genes may be rapidly and efficiently assembled in a specific order and orientation relative to each other.

In still another embodiment of the method of the invention, partial or complete eukaryotic or prokaryotic genomes may be both assembled and modified simultaneously. The method of the invention enables users to alter or mutagenize one or more of the genes or gene fragments, resulting in the creation of genetic alterations such as a mutated gene, a gene deletion, an enhanced gene function, a fusion gene, an altered regulation of the gene functionality, an addition or deletion of restriction enzyme sites or an addition of a site for covalent or non-covalent attachment of a biological or chemical molecule ("handle").

Viral genomic libraries can be created, for example, for the following viruses: (a) all bacteriophage, including all DNA and RNA phage, (b) all eukaryotic viruses, such as baculoviruses and retroviruses, (c) adenoviruses and adeno-associated viruses, Herpes viruses, Vaccinia viruses and all single-stranded, double stranded and partially double stranded DNA viruses, (c) all positive and negative stranded RNA viruses, and (d) replication defective retroviruses. ~~B.~~

ii) Assembly of ~~CDNA~~**cDNA** libraries

Another aspect of the present invention pertains to the assembly of cDNA libraries from individual nucleic acid components. Genes or gene fragments derived from mRNA may be assembled in a manner similar to the above, by synthesizing the resulting cDNA molecules so that they contain unique, and in general, non-palindromic terminal sequences. Such cDNA molecules may then be assembled into eukaryotic or prokaryotic expression vectors. This would allow users to choose from a variety of nucleic acid components derived from cDNA and rapidly and flexibly assemble cDNA libraries . Conventional molecular methods could then be used to select or screen these libraries for the clone or clones of interest.

In the method of the invention, cDNA would be made from mRNA according to art known techniques, described in Maniatis et al., using slight modifications. The method of the present

invention uses modified oligonucleotide primers, containing uracil or alkane diol derivatives as described above, to synthesize a first strand of cDNA resulting in the formation of a unique terminal sequence at the 3' end of the gene. An engineered adaptor, as described above, may be then ligated to the 5' end of a double stranded cDNA molecule, resulting in a unique terminal sequence at the other end of the molecule. The resulting nucleic acid components, encoding the various cDNA molecules, would then be used along with other nucleic acid components encoding appropriate genetic elements, to assemble cDNA library expression vectors.

H. Solid Phase Synthesis

In one embodiment of the method, the nucleic acid components can be linked sequentially to form the nucleic acid construct. This unique attribute lends itself to the automation of construct assembly. The method of the invention uses, preferably, attachment to a solid support as a starting point in the assembly of a series of nucleic acid components, in a defined order, to form a multicomponent nucleic acid construct.

For example, the initial nucleic acid component is attached to a solid support by methods known in the art. Additional nucleic acid components, designed to contain unique terminal sequences at either end, are added in a step-wise fashion, as single components or non-functional multicomponent constructs, and the assembly of components is based on the specific annealing of complementary terminal sequence pairs as previously described. Nucleic acid components may be ligated together, using a ligase enzyme, after each nucleic acid component addition step in the assembly of the larger construct. Unligated DNA fragments may be removed by washing the solid support. Following synthesis, the assembled multicomponent construct or functional construct may be subsequently cleaved from the solid support.

Examples of solid supports that can be used, for the attachment of the initial nucleic acid component, include cellulose, synthetic polymeric material such as modified polystyrenes or polydimethyl acrylamides, and controlled-pore glass. The assembled nucleic acid construct may be cleaved from the solid support by, for example, ammonium hydroxide treatment. Alternatively, the initial nucleic acid component attached to the solid support could be designed to contain a unique restriction site that would be cleaved upon treatment with the appropriate enzyme to release the assembled to nucleic acid construct in solution.~~Kits~~~~The reagents required to practice the method of the invention may be provided in the form of a kit. A kit would comprise, in separate containers, the nucleic acid components to be assembled into a construct, and optionally linking nucleic acid molecules as well as buffers, enzymes and an instructional brochure explaining how to use the kit. In a preferred embodiment the kit would provide the nucleic acid components in an appropriately phosphorylated form for ligation.~~

I. Kits

The reagents required to practice the method of the invention may be provided in the form of a kit. A kit would comprise, in separate containers, the nucleic acid components to be assembled into a construct, and optionally linking nucleic acid molecules as well as buffers, enzymes (e.g., ligase, Klenow, etc)) and an instructional brochure explaining how to use the kit. In a preferred embodiment the kit would provide the nucleic acid components in an

appropriately phosphorylated form for ligation.

The invention further provides a kit for the production of vectors. In one embodiments, the kit for the production of vectors would minimally comprise nucleic acid components encoding origins of replication and selectable markers and optionally, transcriptional regulatory sequence(s). The kit could also include nucleic acid components encoding other vector functions (e.g. a promoter, a transcription or translation regulatory element, etc.).

J. Illustrative Applications Employing the Constructs of the Invention

The nucleic acid constructs produced by the method of the invention, can be employed in an application selected from the group consisting of prokaryotic, eukaryotic (mammalian or non-mammalian) expression. For example, the expression vectors can be used for expression in bacterial cells such as E. coli, insect cells (using baculovirus expression vectors) yeast cells or mammalian cells or they can be transcribed and translated in vitro, for example using the T7 promoter regulatory sequences and the T7 polymerase.. Alternatively, the nucleic acid constructs can be employed in the construction of unique cDNA libraries, protein, antibody and peptide phage display libraries. Kits for screening phage display libraries are commercially available (e.g., the Stratagene SurfZAP.TM. Phage Display Kit, Catalog No. 240612). The constructs can further be employed in gene transfer, gene therapy, and the creation of transgenic organisms, as described above. Finally, vector constructs assembled using the method of the invention may also be used as templates to synthesize RNA using standard methods. Examples of RNA molecules which could be made, would include, without limitation, the following: mRNA, tRNA, rRNA, snRNA, hnRNA, viral or phage RNA, or modified RNA genes or genetic elements.

i) GEOS optimization of gene therapy vectors

DNA-based gene therapy techniques require efficient import of the therapeutic DNA construct into the target cell nucleus where expression occurs. In one aspect, the subject method can be utilized for optimization of a gene therapy vector, e.g., in order to optimize such features as selectivity (tropism), control of gene expression, immunogenicity, and the like, as well as for optimization of a therapeutic protein or nucleic acid expressed by such gene therapy constructs. In this manner, the GEOS method can also be used for the development of safe viral vectors, e.g., is to prevent the generation of replication-competent virus during vector production in a packaging cell line or during gene therapy treatment of an individual.

In one embodiment, the subject GEOS method is applied to the optimization of a gene therapy vector based on a retrovirus, an adenovirus, an adeno-associated virus, a Herpes virus, an HIV virus or other lentivirus, and the like.

In an illustrative embodiment, the subject method is used to optimize the specifity (e.g., infectivity and/or gene expression) of an adenoviral vector for a particular tissue type, such as smooth muscle cells. Adenovirus is a nuclear DNA virus with a genome of about 36 kb, which has been well-characterized through studies in classical genetics and molecular

biology (Horwitz, M. S., "Adenoviridae and Their Replication," in Virology, 2nd edition, Fields, B. N., et al., eds., Raven Press, New York, 1990). The genome is classified into early (known as E1-E4) and late (known as L1-L5) transcriptional units, referring to the generation of two temporal classes of viral proteins. The demarcation between these events is viral DNA replication.

Adenovirus-based vectors offer several unique advantages, including tropism for both dividing and non-dividing cells, minimal pathogenic potential, ability to replicate to high titer for preparation of vector stocks, and the potential to carry large inserts (Berkner, K. L., Curr. Top. Micro. Immunol. 158:39-66, 1992; Jolly, D., Cancer Gene Therapy 1:51-64, 1994). The cloning capacity of an adenovirus vector is about 8 kb, resulting from the deletion of certain regions of the virus genome dispensable for virus growth, e.g., E3, deletions of regions whose function is restored in trans from a packaging cell line, e.g., E1, and its complementation by 293 cells (Graham, F. L., J. Gen. Virol. 36:59-72, 1977), as well as the upper limit for optimal packaging which is about 105%-108% of wild-type length.

The GEOS system can be used to generate a combinatorial library of adenoviral vectors with the goal of selecting those variants which selectively infect smooth muscle cells and/or selectively express a recombinant gene in smooth muscle cells. Such vector libraries can be derived by the GEOS method by combination of libraries of such nucleic acid components as (i) transcriptional regulatory sequence, which may be further broken down into sub-categories of promoters and enhancers, (ii) variants of viral early genes such as E1, E2, E3, E4, and MLP-L1, including loss-of-function mutants and deletions, and (iii) other adenoviral sequences which give rise to the various Ad subtypes (over 40 adenoviral subtypes have been isolated from humans).

In other embodiments, the subjet GEOS method is used to optimize the therapeutic gene which is to be delivered by the gene therapy vector. For instance, secretion or cellular localization of a therapeutic protein can optimized, as appropriate, by the subject method.

ii) GEOS recombination of protein domains

In one aspect, a goal of the present combinatorial method is to increase the number of novel genes and gene products that can be created by "domain shuffling" in a reasonable period of time. As described herein, polypeptide domains can be a polypeptide sequences derived from naturally occurring proteins, or can be artificial in sequence. In certain embodiments, the domain can be a nucleic acid sequences of other function, such as a sequence derived from a ribozyme. By accelerated molecular evolution through shuffling of such domains, a far greater population of novel gene products can be generated and screened in a meaningful period of time.

In one embodiment, the field of application of the present combinatorial method is in the generation of novel enzymatic activities, such as proteolytic enzymes. For example, combinatorial domain-shuffling can be used to rapidly generate a library of potential thrombolytic agents by randomly shuffling the domains of several known blood serum proteins. In another embodiment, the domain-shuffling technique can be used to generate a

library of antibodies from which antibodies of particular affinity for a given antigen can be isolated. As described below, such an application can also be especially useful in grafting CDRs from one variable region to another, as required in the "humanization" of non-human antibodies. Similarly, the present technology can be extended to the immunoglobulin-super family, including the T-cell receptor, etc., to generate novel immulogically active proteins.

In another illustrative embodiment, the present domain-shuffling method can be used to generate novel signal-transduction proteins which can subsequently be used to generate cells which have altered responses to certain biological ligands or stimuli. For instance, protein tyrosine kinases play an important role in the control of cell growth and differentiation. Ligand binding to the extracellular domain of receptor tyrosine kinases often provides an important regulatory step which determines the selectivity of intracellular signaling pathways. Combinatorial domain-shuffling can be used to shuffle, for example, intracellular domains of receptor molecules or signal transduction proteins, including SH2 domains, SH3 domains, kinase domains, phosphatase domains, and phospholipase domains. In another embodiment, variant of SH2 and SH3 domains are randomly shuffled with domains engineered as either protein kinase or phosphatase inhibitors and the combinatorial polypeptide library screened for the ability to block the function of, for example, the action of oncogenic proteins such as sic or ras.

Many techniques are known in the art for screening gene products of combinatorial libraries made by point mutations, and for screening cDNA libraries for gene products having a certain property. Such techniques will be generally applicable to screening the gene libraries generated by the present domain-shuffling methodology. The most widely used techniques for screening large gene libraries typically comprises cloning the gene library into replicable expression vectors, transformning appropriate cells with the resulting library of vectors, and expressing the combinatorial genes under conditions in which detection of a desired activity facilitates relatively easy isolation of the vector encoding the gene whose producted was detected. For instance, in the case of shuffling intracellular domains, phenstypic changes can be detected and used to isolate cells expressing a combinatorially-derived gene product conferring the new phenotype. Likewise, interaction trap assays can be used in vivo to screen large polypeptide libraries for proteins able to bind a "bait" protein, or alternatively, to inhibit binding of two proteins.

The domain shuffling methods described herein can be used to create new ribozymes. For ribozymes, one illustrative embodiment comprises screening a ribozyme library for the ability of molecules to cleave an mRNA molecule and disrupt expression of a protein in such a manner as to confer some phenotypic change to the cell.

In another embodiment, the subject GEOS method can be used to generate libraries of composite transcription factors, e.g., which are chimeric combinations of DNA binding domains and activation domains. Such composite factors can be optimized for, e.g., conditional sensitivity (inducibility or repressibility), level of expression when activated, cell-type specifity of expression, recognition of unique transcriptional regulatory elements,

recruitment of basal and non-basal transcriptional complexes, and the like. Such composite factors may be useful in gene therapy, plant engineering, recombinaint protein production and general research.

Nuclear localization signals, for example, have been shown to facilitate efficient nuclear localization of macromolecules. In order to facilitate nuclear import of an exogenous therapeutic DNA construct, peptide nuclear localization signals can be joined to a GEOS vector assembly by utilizing techniques for the covalent attachment of cationic peptides to double stranded DNA with a chemical cross-linker, resulting in a significant increase in nuclear uptake of the resulting construct (see e.g. Wolff, et al. (1998) Biotechnology 16: 80-85). Indeed, solid phase synthesis of a gene therapeutic construct using GEOS methodology could be directly followed with coupling to an appropriate cationic peptide nuclear localization signal allowing direct synthesis of the DNA therapeutic agent without an intermediate amplification step.

In yet another screening assay, the gene product, especially if its a polypeptide, is displayed on the surface of a cell or viral particle, and the ability of particular cells or viral particles to bind another molecule via this gene product is detected in a "panning assay". For example, the gene library can be cloned into the gene for a surface membrane protein of a bacterial cell, and the resulting fusion protein detected on the surface of the bacteria (Ladner et al., WO 88/06630; Fuchs et al. (1991) Bio/Technology 9:1370-1371; and Goward et al. (1992) TIBS 18:136-140). In another embodiment, gene library is expressed as fusion protein on the surface of a viral particle. For instance, in the filamentous phage system, foreign peptide sequences can be expressed on the surface of infectious phage, thereby conferring two significant benefits. First, since these phage can be applied to affinity matrices at very high concentrations, large number of phage can be screened at one time. Second, since each infectious phage encodes the gene product on its surface, if a particular phage is recovered from an affinity matrix in low yield, the phage can be amplified by another round of infection. The group of almost identical E.coli filamentous phages M13, fd, and f1 are most often used in phage display libraries, as either of the phage gIII or gVIII coat proteins can be used to generate fusion proteins without disrupting the ultimate packaging of the viral particle (Ladner et al. PCT publication WO 90/02909; Garrard et al., PCT publication WO 92/09690; Marks et al. (1992) J. Biol. Chem. 267:16007-16010; Griffths et al. (1993) EMBO J 12:725-734; Clackson et al. (1991) Nature 352:624-628; and Barbas et al. (1992) PNAS 89:4457-4461).

a) Antibody Repertoires

Mouse monoclonal antibodies are readily generated by the fusion of antibody-producing B lymphocytes with myeloma cells. However, for therapeutic applications, human monoclonal antibodies are preferred. Despite extensive efforts, including production of heterohybridomas, Epstein-Barr virus immortalization of human B cells, and "humanization" of mouse antibodies, no general method comparable to the Kohler-Milstein approach has emerged for the generation of human monoclonal antibodies.

Recently, however, techniques have been developed for the generation of antibody libraries

in E. coli capable of expressing the antigen binding portions of immunoglobulin heavy and light chains. For example, recombinant antibodies have been generated in the form of fusion proteins containing membrane proteins such as peptidoglycan-assoicated lipoprotein (PAL), as well as fusion proteins with the capsular proteins of viral particles, or simply as secreted proteins which are able to cross the bacterial membrane after the addition of a bacterial leader sequence at their N-termini. (See, for example, Fuchs et al. (1991) Bio/Technology 9:1370-1372; Bettes et al. (1988) Science 240:1041-1043; Skerra et al. (1988) Science 240:1038-1041; Hay et al. (1992) Hum. Antibod. Hybridomas 3:81-85; and Barbas et al. International Publication No. WO92/18019).

The display of antibody fragments on the surface of filamentous phage that encode the antibody gene, and the selection of phage binding to a particular antigen, offer a powerful means of generating specific antibodies in vitro. Typically, phage antibodies (phAbs) have been generated and expressed in bacteria by cloning repertoires of rearranged heavy and light chain V-genes into filamentous bacteriophage. Antibodies of a particular specificity can be selected from the phAb library by panning with antigen.

The present combinatorial approach can be applied advantageously to the production of recombinant antibodies by providing antibody libraries not readily accessible by any prior technique. For instance, in contrast to merely sampling combinations of V.sub.H and V.sub.L chains, the present method allows the complementarity-determining regions (CDRs) and framework regions (FRs) themselves to be randomly shuffled in order to create novel V.sub.H and V.sub.L regions which were not represented in the originally cloned rearranged V-genes.

Antibody variable domains consist of a .beta.-sheet framework with three loops of hypervariable sequences (e.g. the CDRs), and the antigen binding site is shaped by loops from both heavy (V.sub.H) and light (V.sub.L) domains. The loops create antigen binding sites of a variety of shapes, ranging from flat surfaces to pockets. For human V.sub.H domains, the sequence diversity of the first two CDRs are encoded by a repertoire of about 50 germline V.sub.H segments (Tomlinson et al. (1992) J. Mol. Biol. 227:). The third CDR is generated from the combination of these segments with about 30 D and six J segments (Ichihara et al. (1988) EMBO J 7: 4141-4150). The lengths of the first two CDRs are restricted, with the length being 6 amino acid residues for CDR1, 17 residues, and for CDR2. However, the length of CDR3 can differ significantly, with lengths ranging from 4 to 25 residues.

For human light chain variable domains, the sequence diversity of the first two CDRs and part of CDR3 are encoded by a repertoire of about 50 human V.sub.kappa. segments (Meindl et al. (1990) Eur. J. Immunol. 20: 1855-1863) and >10 V.sub.lambda. segments (Chuchana et al. (1990) Eur. J. Immunol. 20: 1317-1325; and Combriato et al. (1991) Eur. J. Immunol. 21: 1513-1522). The lengths of the CDRs are as follows, CDR1=11-14 residues; CDR2=8 residues; and CDR3 ranges from 6 to 10 residues for V.sub.kappa. genes and 9 to 13 for V.sub.lambda. genes.

The present invention contemplates combinatorial methods for generating diverse antibody

libraries, as well as reagents and kits for carrying out such methods. In one embodiment, the present combinatorial approach can be used to recombine both the framework regions and CDRs to generate a library of novel heavy and light chains. In another embodiment, domain-shuffling can be used to shuffle only the framework regions which flank specific CDR sequences. While both schemes can be used to generate antibodies directed to a certain antigen, the later strategy is particularly amenable to being used for "humanizing" non-human monoclonal antibodies.

In one embodiment, the combinatorial units useful for generating diverse antibody repertoires by the present domain-shuffling methods comprise exon constructs corresponding to fragments of various immunoglobulin variable regions flanked by intronic sequences that can drive their ligation. For example, the "exonic" sequences of the combinatorial units can be selected to encode essentially just a framework region or CDR; or can be generated to correspond to larger fragments which may include both CDR and FR sequences. The combinatorial units can be made by standard cloning techniques that manipulate DNA sequences into vectors which provide appropriate flanking intron fragments with non-palindromic overhangs.

Methods are generally known for directly obtaining the DNA sequence of the variable regions of any immunoglobulin chain by using a mixture of oligomer primers and PCR. For instance, mixed oligonucleotide primers corresponding to the 5' leader (signal peptide) sequences and/or FR1 sequences and a conserved 3' constant region primer have been used for PCR amplification of the heavy and light chain variable regions from a number of human antibodies directed to, for example, epitopes on HIV-1 (gp 120, gp 42), digoxin, tetanus, immunoglobulins (rheumatoid factor), and MHC class I and II proteins (Larrick et al. (1991) Methods: Companion to Methods in Enzymology 2: 106-110). A similar strategy has also been used to amplify mouse heavy and light chain variable regions from murine antibodies, such as antibodies raised against human T cell antigens (CD3, CD6), carcino embryonic antigen, and fibrin (Larrick et al. (1991) Bio Techniques 11: 152-156).

In the present invention, RNA is isolated from mature B cells of, for example, peripheral blood cells, bone marrow, or spleen preparations, using standard protocols. First-strand cDNA is synthesized using primers specific for the constant region of the heavy chain(s) and each of the .kappa. and .lambda. light chains. Using variable region PCR primers, such as those shown in Table II below, the variable regions of both heavy and light chains are amplified (preferably in separate reactions) and ligated into appropriate expression vectors. The resulting libraries of vectors (e.g. one for each of the heavy and light chains) contain a variegated population of variable regions. The intronic addition can be carried out simultaneously for all three FR/CDR boundaries, or at fewer than all three boundaries. So, for example, leader-FR1 (IVS 1-4), (IVS 5,6)CDR1,FR2(IVS 1-4), (IVS 5,6)CDR2,FR3(IVS 1-4) and (IVS 5,6)CDR3,FR4 combinatorial units can be generated with flanking non-palindromic overhangs.

~~The invention further provides a kit for the production of vectors. The kit for the production of vectors would minimally comprise nucleic acid components encoding origins of replication, selectable markers and inserts of interest. The kit could also include nucleic acid components~~

encoding other vector functions (e.g. a promoter, a transcription or translation regulatory element, etc.).

Applications Employing the Constructs of the Invention

The nucleic acid constructs produced by the method of the invention, can be employed in an application selected from the group consisting of prokaryotic, eukaryotic (mammalian or non-mammalian) expression. For example, the expression vectors can be used for expression in bacterial cells such as E. coli, insect cells (using baculovirus expression vectors) yeast cells or mammalian cells or they can be transcribed and translated in vitro, for example using the T7 promoter regulatory sequences and the T7 polymerase. Alternatively, the nucleic acid constructs can be employed in the construction of unique cDNA libraries, protein, antibody and peptide phage display libraries. Kits for screening phage display libraries are commercially available (e.g., the Stratagene SurfZAP.TM. Phage Display Kit, Catalog No. 240612). The constructs can further be employed in gene transfer, gene therapy, and the creation of transgenic organisms, as described above. Finally, vector constructs assembled using the method of the invention may also be used as templates to synthesize RNA using standard methods. Examples of RNA molecules which could be made, would include, without limitation, the following: MRNA, tRNA, rRNA, snRNA, hnRNA, viral or phage RNA, or modified RNA genes or genetic elements.

EXAMPLES

TABLE II

Human Immunoglobulin Variable Region PCR Primers

5' End Sense

Human heavy chains

Group A

5'-GGGAATTCATGGACTGGACCTGGAGG(AG)TC(CT)-
TCT(GT)C-3'

Group B

5'-GGGAATTCATGGAG(CT)TTGGGCTGA(CG)CTGG(CG)-
TTTT-3'

Group C

5'-GGGAATTCATG(AG)A(AC)(AC)(AT)ACT(GT)TG(GT)-
(AT)(CG)C(AT)(CT)(CG)CT(CT)CTG-3'

Human .kappa. light chain

5'-GGGAATTCATGGACATG(AG)(AG)(AG)(AGT)(CT)CC-
(ACT)(ACG)G(CT)GT)CA(CG)CTT-3'

Human .lambda. light chain

5'-GGGAATTCATG(AG)CCTG(CG)(AT)C(CT)CCTCTC(CT)-
T(CT)CT(CG)(AT)(CT)C-3'

3' End sense constant region

Human IgM heavy chain

5'-CCAAGCTTAGACGAGGGGGAAAAGGGTT-3'

Human IgG1 heavy chain

5'-CCAAGCTTGGAGGAGGGTGCCAGGGGG-3'

Human .lambda. light chain

5'-CCAAGCTTGAAGCTCCTCAGAGGAGGG-3'

Human .kappa. light chain

5'-CCAAGCTTTCATCAGATGGCGGGAAGAT-3'

Murine Immunoglobulin Variable Region PCR Primers

5' End sense

Leader (signal peptide) region
(amino acids -20 to -13)
Group A
5'-GGGGAATTCATG(GA)A(GC)TT(GC)(TG)GG(TC)T(AC)-
A(AG)CT(GT)G(GA)TT-3'
Group B
5'-GGGGAATTCATG(GA)AATG(GC)A(GC)CTGGGT(CT)-
(TA)T(TC)CTCT-3'
Framework 1 region (amino acids 1 to 8)
5'-GGGGAATTC(CG)AGGTG(CA)AGCTC(CG)(AT)(AG)(CG)-
A(AG)(CT)C(CG)GGG-3'
3' End sense constant region
Mouse .gamma. constant region (amino acids 121 to 131)
5'-GGAAGCTTA(TC)CTCCACACACAGG(AG)(AG)CCAGTG-
GATAGAC-3'
Mouse .kappa. light chain (amino acids 116 to 122)
5'-GGAAGCTTACTGGATGGTGGGAAGATGGA-3'
Bases in parentheses represent substitutions at a given residue. EcoRI and
HindIII sites are underlined.

The following examples are by way of illustration and are not intended to limit the claims. Persons of skill will readily recognize that the protocols of the examples can be modified in numerous non-critical ways.

Example 1Simultaneous assembly of a viable plasmid vectorTo demonstrate the simultaneous assembly of multiple nucleic acid components having unique, non-palindromic terminal sequences, to produce a viable plasmid vector, three nucleic acid components are used. The first nucleic acid component is a gene coding for green fluorescent protein, 0.7 Kb in length, the second one is a 0.6 Kb molecule coding for terminator sequences and a histidine tag, and the third one is a 2.5 Kb molecule coding for the lac promoter, an ampicillin resistance gene, and an origin of replication.1. Synthesis of the Nucleic Acid ComponentsThe nucleic acid components used in the present example are synthesized by PCR amplification. The PCR reactions are performed in varying volumes (in general, 10-100 microliters) containing a 50 mM KC1, 10 mM Tris HCl (pH 8.4), 1.5 mM MgCl.sub.2 buffer and 0.2 mM of each dNTP, 1.25 units of taq DNA polymerase, 10.sup. 5 M template molecules, and 20 pmol of each primer. The primers used contain uracil residues at specific locations in order to generate 3' terminal sequences as described in U.S. Pat. No. 5,137,814. The PCR reaction is carried out using a thermal cycling instrument, where there is an initial denaturation phase of 95.degree. C. for 5 minutes, followed by multiple cycles (20-40 cycles) of a denaturation step at 94.degree. C., an annealing step at 37.degree. 65.degree. C. and an extension step at 72.degree. C. The resulting PCR products are analyzed by gel electrophoresis to determine size and purity.2. Generation of Terminal sequencesFollowing PCR amplification and purification of the correct size fragments, the PCR products (approximately 100-200 ng) are dissolved in 10 microliters of the UDG reaction buffer (25 mM Tris-HCl (pH 7.8), 10 mM Mg.sub.2 Cl, 4 mM beta-mercaptoethanol, 0.4 mM ATP). Single-stranded 3' Terminal sequences are made by treatment of the PCR product with 1-2 units of uracil DNA glycosidase (UDG) for 10 minutes at 37.degree. C. The enzyme is inactivated and reaction is terminated by heating the sample at 65.degree. C. for 10 minutes.3. Assembly and Ligation of the Nucleic Acid ComponentsTo assemble the vector the individual purified nucleic acid components are mixed in equimolar amounts (approximately 20-200 ng total in 20 microliters) in the UDG treatment buffer and heated to 65.degree. C., followed by gradually

~~cooling down to room temperature (25.degree. C.), to permit efficient annealing of the complementary ends of the nucleic acid components. The reaction mixture may optionally be treated with T4 DNA ligase at 14.degree. C. overnight to ligate the nucleic acid components or used directly to transform competent bacterial hosts.4. TransformationA 10 .mu.l aliquot of the assembled vector is added to 100 .mu.l of competent E. coli cells (DH5.alpha.), transformed following the manufacturers recommendations, and plated on LB plates containing ampicillin and IPTG.5. Analysis of the Vector ConstructIsolated fluorescent colonies are selected and pure DNA plasmid prepared using a mini-prep. Correct assembly of the vector construct is determined using standard molecular biological methods, such as restriction enzyme digestion and agarose gel electrophoresis.EquivalentsThose skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.~~

Optionally, the leader-FR1 (IVS 1-4) construct can be linked to an insoluble resin by standard techniques, and each set of combinatorial units (CDR1/FR2, CDR2/FR3, CDR3/FR4) can be sequentially annealed to the resin-bound nucleic acid with unbound reactants washed away between each round of addition. After addition of the (IVS 5,6)CDR3,FR4 units to the resin bound molecules, the resulting trans-spliced molecule can be released from the resin, PCR amplified using primers for the leader sequence and constant region, and subsequently cloned into an appropriate vector for generating a screenable population of antibody molecules.

Taking the dissection of the variable regions one step further, a set of domain libraries can be generated for ordered combinatorial ligation much the same as above, except that each combinatorial unit is flanked at its 5' end with a non-palindromic overhang sequence that is unable to drive a domain-shuffling reaction with the non-palindromic overhang sequence at its 3' end. With regard to ordered gene assembly, each combinatorial unit is effectively protected from addition by another unit having identical flanking intron fragments.

Furthermore, CDR combinatorial units can be generated which are completely random in sequence, rather than cloned from any antibody source. For example, a degenerate oligonucleotide can be synthesized for CDR1 which encodes all possible amino acid combinations for the 6 a.a. sequence. The nucleotide sequences which flank the CDR-encoding portion of the oligonucleotide comprise the flanking intron sequences and non-palindromic overhang sequences necessary to allow annealing of the degenerate oligonucleotide into the plasmid and reconstitute a construct which would produce a spliceable transcript. To avoid creation of stop codons which can result when codons are randomly synthesized using nucleotide monomers, "dirty bottle" synthesis can instead be carried out using a set of nucleotide trimers which encode all 20 amino acids.

With slight modification, the present ordered combinatorial ligation can be used to efficiently humanize monoclonal antibodies of non-human origin. The CDRs from the monoclonal antibody can be recombined with human framework region libraries (e.g. an FR1 library, an FR2 library, etc.) to produce a combinatorial population of variable regions in which the CDR sequences are held constant, but each of the framework regions

have been randomized. The variable regions can be subsequently fused with sequences corresponding to the appropriate human constant regions, and the antibodies resulting from heavy and light chain association can be screened for antigen binding using standard panning assays such as phage display. In contrast to contemporary humanization schemes which require the practitioner to prejudicially choose a particular human scaffold into which the CDRs are grafted, the present technique provides a greater flexibility in choosing appropriate human framework regions which do not adversely affect antigen binding by the resultant chimeric antibody.

To illustrate, the variable regions of both the heavy and light chains of a mouse monoclonal antibody can be cloned using primers as described above. The sequence of each CDR can be obtained by standard techniques. The CDRs can be cloned into vectors which provide appropriate flanking intronic sequences and non-palindromic overhang sequences. As described above, the particular intronic fragments provided with each murine CDR and each human FR construct can be selected to disfavor multiple ligations at each step of addition to a resin bound nucleic acid. The library of human heavy chain leader, FR1(IVS-1-4) constructs can be immobilized on a resin, and in a first round of ligation, the heavy chain murine (IVS-5,6) CDR1 (IVS-1-4) construct is added under conditions which facilitate annealing of the overhang sequences. Un-ligated combinatorial units are washed away, and the library of human heavy chain (IVS-5,6) FR2 (IVS-1-4) units are admixed and ligated to the resin-bound nucleic acids terminating with the murine CDR construct. This process is carried out for the remaining murine CDR and human FR units of the heavy chain, and a similar process is used to construct combinatorial light chain chimeras as well. The resulting chimeric heavy and light chains can be cloned into a phage display library, and the phAbs screened in a panning assay to isolate humanized antibodies (and their genes) which bind the antigen of interest.

b) Combinatorial Enzyme Libraries

The subject method can also be used to generate novel enzymatic activities. In one embodiment, the subject combinatorial method can be used to generate novel blood-clotting or anticoagulant enzymes. Plasminogen activators (PAs) are a class of serine proteases that convert the proenzyme plasminogen into plasmin, which then degrades the fibrin network of blood clots. The plasminogen activators have been classified into two immunologically unrelated groups, the urokinase-type PAs (u-PA) and the tissue-type PA (tPA), with the later activator being the physiological vascular activator. These proteins, as well as other proteases of the fibrinolytic pathway, are composed of multiple structural domains which appear to have evolved by genetic assembly of individual domains with specific structural and/or functional properties. For instance, the amino terminal region of tPA is composed of multiple structural/functional domains found in other plasma proteins, including a "finger-like domain" homologous to the finger domains of fibronectin, an "epidermal growth factor domain" homologous to human EGF, and two disulfide-bonded triple loop structures, commonly referred to as "kringle domains", homologous to the kringle regions in plasminogen. The region comprising residues 276-527 (the "catalytic domain" is homologous to that of other serine proteases and contains the catalytic triad. In addition, the gene for tPA encodes a signal secretion peptide which directs secretion of the

protein into the extracellular environment, as well as a pro-sequence which is cleaved from the inactive form of the protease (the "plasminogen") to active tPA during the fibrinolytic cascade.

These distinct domains in tPA are involved in several functions of the enzyme, including its binding to fibrin, stimulation of plasminogen activation by fibrin, and rapid in vivo clearance. Approaches used to characterize the functional contribution of these structural domains include isolation of independent structural domains as well as the production of variant proteins which lack one or more domains. For example, the fibrin selectivity of tPA is found to be mediated by its affinity for fibrin conferred by the finger-like domain and by at least one of the kringle domains.

The present combinatorial method can be used to generate novel plasminogen activators having superior thrombolytic properties, by generating a library of proteins through shuffling of the domains of plasma proteins. As described below, one mode of generating the combinatorial library comprises the random domain-shuffling of a mixture of coding sequences corresponding to each of the domains of the mature tPA protein. Briefly, a cDNA clone of tPA is obtained and, through the use of specific PCR amplimers, each of the 5 protein domains is amplified and isolated. Each of these amplified domains is then separately cloned into a plasmid as an exon module such that the 5' end of the exon is preceded by group II domains V-VI, and the 3' end of the exon is followed by group II domains I-III. Generation of style-stranded non-palindromic overhangs, and mixture of these constructs under annealing conditions, can result in random ligation of the exons to one and other and assembly of the combinatorial gene library which can subsequently be screened for fibrinolytic activity.

Moreover, combinatorial units can be generated from other proteins, including proteins having no catalytic role in blood clotting or fibrinolysis. For example, a library of catalytic domains can be generated from other thrombolytic proteases, blood clotting factors, and other proteases having peptidic activity similar to the typsin-like activity of tPA. Likewise, libraries of splicing constructs can be derived from EGF-like domains, finger-like domains, kringle domains, and Calcium-binding domains from a vast array of proteins which contain such moieties.

iii) GEOS-mediated generation of nucleic acid vaccines

Modem vaccine technnology allows for the stimulation of antibodies to a given antigen by direct administration of a nucleic acid construct encoding that antigen. The GEOS methodology is particularly suited to this application. In this application, the GEOS insert nucleic acid component corresponds to the antigen to which immunity is directed. There are at least three types of methods for producing DNA vaccines: DNA vaccines consisting of E. Coli--derived expression vectors encoding the antigen of interest; recombinant immunoglobulin molecules containing foreign epitopes grafted into complementarity-determining regions (CDRs) resulting in the formation of "antigenized antibodies" which can induce immune responses against these engineered epitopes; and a merging of the two approaches called somatic transgene immunization (STI--see, Zanetti, et al. (1997) Nature

Biotechnology, 15: 876-86) whereby an immunoglobulin heavy chain containing heterologous antigenic epitopes engineered into one or more CDRs, followed by the interspleenic innoculation of the DNA construct using tissue specific promoter and enhancer elements.

Construction of DNA vaccine constructs is particularly suited to the method of the present invention due to the flexibility of being able to interchange specific components in the discovery process. For example, when applied to the somatic transgene immunization method, the GEOS methodology allows for the rapid assembly and systematic variation of the critical STI DNA vaccine vectors components. These include: a general component for the expression vector backbone, components corresponding to frame-work regions (FRs), components corresponding to complementarity-determining regions (CDRs), components corresponding to tissue-specific immunoglobulin promoter and/or enhancer elements, and one or more antigenic epitope components to be inserted into any one or more of the CDR domains. A particular advantage of the invention in this application is the ability to vary: the CDR domain into which the immunogenic peptide fragment is to be substituted (i.e. CDR1, 2 or 3), the immunoglobulin chain into which the immunogenic peptide fragment is to be substituted (light chain or heavy chain), the number of CDR/ immunogenic peptide fragment substitutions (from 1 to 3 per chain), and the tissue tropism of the resulting construct (through variation of the enhancer element component). Another advantage is that individual immunogenic peptide fragment elements can be assembled into the vector system as heterogeneous pools corresponding to different but related epitope families. Such "pooled vaccines" find particular utility in stimulating antibodies which recognize highly variable regions of, for example, viral coat proteins.

GEOS components could also assist in immunoglobulin scaffold experiments where combinatorial arrays of scaffold sequences comprise at least one of the nucleic acid components. The ability to carry out multiplex experiments in which candidate scaffold components are combined with selected epitope candidates would allow for the generation of very large numbers of constructs from relatively few components. Furthermore the flexibility in vector assembly afforded by the GEOS methodology would allow one to switch rapidly between a phage display immunoglobulin system (which can undergo rapid selection processes for high-affinity epitope recognition) and a corresponding GEOS DNA vaccine construct.

iv) Two-hybrid systems

The various yeast and mammalian two-hybrid systems allow for the rapid cloning and characterization of interacting polypeptides which are expressed as "bait" and "prey" constructs. The bait construct typically comprises a DNA binding domain element and a first "gene of interest" element which are fused in-frame. The prey construct typically comprises a transcriptional activation domain and a second "gene of interest" element which are fused in-frame. "Global grids" corresponding to all possible combination of multiple genes of interest can be used to investigate the association of various proteins in a biological system. The two-hybrid system allows the investigator to determine which genes encode proteins which interact with one another. Although standard cloning techniques are

suitable for two-hybrid analysis of relatively small genetic systems, they are not efficient for the analysis of large genetic systems. The GEOS methodology, through delivery of heterogeneous pools comprising multiple "gene of interest" nucleic acid components, is particularly well suited to the systematic synthesis of two-hybrid bait and prey vectors from large numbers of pooled cDNAs. Such a pooled vector assemble strategy would be particularly useful in analyzing the huge number of novel cDNAs being recovered in human and mouse genome sequencing efforts. The GEOS vector assembly methodology provides a facile means for characterizing large numbers of cDNAs for potential biological interaction of their encoded products, thereby providing an initial biological characterization of these genes into related, biologically-interacting families.

v) Biosynthesis

Metabolic engineering can be used to alter organisms to optimize the production of practically any metabolic intermediate, including antibiotics, vitamins, amino acids such as phenylalanine and aromatic amino acids, ethanol, butanol, polymers such as xanthan gum and bacterial cellulose, peptides, and lipids. When such compounds are already produced by a host, the subject recombination techniques described herein can be used to optimize production of the desired metabolic intermediate, including such features as increasing enzyme substrate specificity and turnover number, altering metabolic fluxes to reduce the concentrations of toxic substrates or intermediates, increasing resistance of the host to such toxic compounds, eliminating, reducing or altering the need for inducers of gene expression/activity, increasing the production of enzymes necessary for metabolism, etc.

Enzymes can also be evolved for improved activity in solvents other than water. This is useful because intermediates in chemical syntheses are often protected by blocking groups which dramatically affect the solubility of the compound in aqueous solvents. Many compounds can be produced by a combination of pure chemical and enzymically catalyzed reactions. Performing enzymic reactions on almost insoluble substrates is clearly very inefficient, so the availability of enzymes that are active in other solvents will be of great use. One example of such a scheme is the evolution of a paranitrobenzyl esterase to remove protecting groups from an intermediate in loracarbef synthesis (Moore, J. C. and Arnold, F. H. Nature Biotechnology 14:458-467 (1996)). In this case alternating rounds of error-prone PCR and colony screening for production of a fluorescent reporter from a substrate analogue were used to generate a mutant esterase that was 16-fold more active than the parent molecule in 30% dimethylformamide. No individual mutation was found to contribute more than a 2-fold increase in activity, but it was the combination of a number of mutations which led to the overall increase. Structural analysis of the mutant protein showed that the amino acid changes were distributed throughout the length of the protein in a manner that could not have been rationally predicted. Sequential rounds of error-prone PCR have the problem that after each round all but one mutant is discarded, with a concomitant loss of information contained in all the other beneficial mutations. The subject GEOS recombination method avoids this problem, and would thus be ideally suited to evolving enzymes for catalysis in other solvents, as well as in conditions where salt concentrations or pH were different from the original enzyme optimas.

In addition, the yield of almost any metabolic pathway can be increased, whether consisting entirely of genes endogenous to the host organisms or all or partly heterologous genes. Optimization of the expression levels of the enzymes in a pathway is more complex than simply maximizing expression. In some cases regulation, rather than constitutive expression of an enzyme may be advantageous for cell growth and therefore for product yield, as seen for production of phenylalanine (Backman et al. Ann. NY Acad. Sci. 589:16-24 (1990)) and 2-keto-L-gluconic acid (Anderson et al. U.S. Pat. No. 5,032,514). In addition, it is often advantageous for industrial purposes to express proteins in organisms other than their original hosts. New host strains may be preferable for a variety of reasons, including ease of cloning and transformation, pathogenicity, ability to survive in particular environments and a knowledge of the physiology and genetics of the organisms. However, proteins expressed in heterologous organisms often show markedly reduced activity for a variety of reasons including inability to fold properly in the new host (Sarthy et al. Appl. Environ. Micro. 53:1996-2000 (1987)). Such difficulties can indeed be overcome by the recombination strategies of the instant invention.

a. Antibiotics

The range of natural small molecule antibiotics includes but is not limited to peptides, peptidolactones, thiopeptides, beta-lactams, glycopeptides, lantibiotics, microcins, polyketide-derived antibiotics (anthracyclins, tetracyclins, macrolides, avermectins, polyethers and ansamycins), chloramphenicol, aminoglycosides, aminocyclitols, polyoxins, agrocins and isoprenoids.

There are at least three ways in which the GEOS recombination techniques of the instant invention can be used to facilitate novel drug synthesis, or to improve biosynthesis of existing antibiotics.

First, antibiotic synthesis enzymes can be "evolved" together with transport systems that allow entry of compounds used as antibiotic precursors to improve uptake and incorporation of function-altering artificial side chain precursors. For example, penicillin V is produced by feeding Penicillium the artificial side chain precursor phenoxyacetic acid, and LY146032 by feeding Streptomyces roseosporus decanoic acid (Hopwood, Phil. Trans. R. Soc. Lond. B 324:549-562 (1989)). Poor precursor uptake and poor incorporation by the synthesizing enzyme often lead to inefficient formation of the desired product. The use of the subject recombination method on these two systems can increase the yield of desired product.

Furthermore, a combinatorial approach can be taken in which an enzyme is shuffled for novel catalytic activity/substrate recognition (perhaps by including randomizing oligonucleotides in key positions such as the active site). A number of different substrates (for example, analogues of side chains that are normally incorporated into the antibiotic) can then be tested in combination with all the different enzymes and tested for biological activity. In this embodiment, plates are made containing different potential antibiotic precursors (such as the side chain analogues). The microorganisms containing the shuffled library (the library strain) are replicated onto those plates, together with a competing,

antibiotic sensitive, microorganism (the indicator strain). Library cells that are able to incorporate the new side chain to produce an effective antibiotic will thus be able to compete with the indicator strain, and will be selected for.

Second, the expression of heterologous genes transferred from one antibiotic synthesizing organism to another can be optimized. The newly introduced enzyme(s) act on secondary metabolites in the host cell, transforming them into new compounds with novel properties. Using traditional methods, introduction of foreign genes into antibiotic synthesizing hosts has already resulted in the production of novel hybrid antibiotics. Examples include mederrhodin, dihydrogranatirhodin, 6-deoxyerythromycin A, isovalerylspiramycin and other hybrid macrolides (Cameron et. al. Appl. Biochem. Biotechnol. 38:105-140 (1993)). The GEOS recombination techniques of the instant invention can be used to optimize expression of the foreign genes, to stabilize the enzyme in the new host cell, and to increase the activity of the introduced enzyme against its new substrates in the new host cell. In some embodiments of the invention, the host genome may also be so optimized.

Third, the substrate specificity of an enzyme involved in secondary metabolism can be altered so that it will act on and modify a new compound or so that its activity is changed and it acts at a different subset of positions of its normal substrate. GEOS recombination can be used to alter the substrate specificities of enzymes. Furthermore, in addition to GEOS recombination of individual enzymes being a strategy to generate novel antibiotics, GEOS recombination of entire pathways, by altering enzyme ratios, will alter metabolite fluxes and may result, not only in increased antibiotic synthesis, but also in the synthesis of different antibiotics. This can be deduced from the observation that expression of different genes from the same cluster in a foreign host leads to different products being formed (see p. 80 in Hutchinson et. al., (1991) Ann NY Acad Sci, 646:78-93). GEOS recombination of the introduced gene clusters may result in a variety of expression levels of different proteins within the cluster (because it produces different combinations of, in this case regulatory, mutations). This in turn may lead to a variety of different end products. Thus, "evolution" of an existing antibiotic synthesizing pathway could be used to generate novel antibiotics either by modifying the rates or substrate specificities of enzymes in that pathway.

Additionally, antibiotics can also be produced in vitro by the action of a purified enzyme on a precursor. For example isopenicillin N synthase catalyses the cyclization of many analogues of its normal substrate (d-(L-a-aminoadipyl)-L-cysteinyl-D-valine) (Hutchinson, Med. Res. Rev. 8:557-567 (1988)). Many of these products are active as antibiotics. A wide variety of substrate analogues can be tested for incorporation by secondary metabolite synthesizing enzymes without concern for the initial efficiency of the reaction. GEOS recombination can be used subsequently to increase the rate of reaction with a promising new substrate.

Thus, organisms already producing a desired antibiotic can be evolved with the GEOS recombination techniques described herein to maximize production of that antibiotic. Additionally, new antibiotics can be evolved by manipulation of genetic material from the host by the GEOS recombination techniques described herein. Genes for antibiotic

production can be transferred to a preferred host after cycles of GEOS recombination. Antibiotic genes are generally clustered and are often positively regulated, making them especially attractive candidates for the GEOS recombination techniques of the instant invention. Additionally, some genes of related pathways show cross-hybridization, making them preferred candidates for the generation of new pathways for new antibiotics by the GEOS recombination techniques of the invention. Furthermore, increases in secondary metabolite production including enhancement of substrate fluxes (by increasing the rate of a rate limiting enzyme, deregulation of the pathway by suppression of negative control elements or over expression of activators and the relief of feedback controls by mutation of the regulated enzyme to a feedback-insensitive deregulated protein) can be achieved by GEOS recombination without exhaustive analysis of the regulatory mechanisms governing expression of the relevant gene clusters.

The host chosen for expression of evolved genes is preferably resistant to the antibiotic produced, although in some instances production methods can be designed so as to sacrifice host cells when the amount of antibiotic produced is commercially significant yet lethal to the host. Similarly, bioreactors can be designed so that the growth medium is continually replenished, thereby "drawing off" antibiotic produced and sparing the lives of the producing cells. Preferably, the mechanism of resistance is not the degradation of the antibiotic produced.

Numerous screening methods for increased antibiotic expression are known in the art, as discussed above, including screening for organisms that are more resistant to the antibiotic that they produce. This may result from linkage between expression of the antibiotic synthesis and antibiotic resistance genes (Chater, Bio/Technology 8:115-121 (1990)). Another screening method is to fuse a reporter gene (e.g. xylE from the Pseudomonas TOL plasmid) to the antibiotic production genes. Antibiotic synthesis gene expression can then be measured by looking for expression of the reporter (e.g. xylE encodes a catechol dioxygenase which produces yellow muconic semialdehyde when colonies are sprayed with catechol (Zukowski et al. Proc. Natl. Acad. Sci. U.S.A. 80:1101-1105 (1983)).

The wide variety of cloned antibiotic genes provides a wealth of starting materials for the GEOS recombination techniques of the instant invention. For example, genes have been cloned from Streptomyces cattleya which direct cephamycin C synthesis in the non-antibiotic producer Streptomyces lividans (Chen et al. Bio/Technology 6;1222-1224 (1988)). Clustered genes for penicillin biosynthesis (delta-(L-alpha-aminoadipyl)-L-cysteinyl-D-valine synthetase; isopenicillin N synthetase and acyl coenzyme A:6-aminopenicillanic acid acyltransferase) have been cloned from Penicillium chrysogenum. Transfer of these genes into Neurospora crassa and Aspergillus niger result in the synthesis of active penicillin V (Smith et al. Bio/Technology 8:39-41 (1990)). For a review of cloned genes involved in Cephalosporin C, Penicillins G and V and Cephamycin C biosynthesis, see Piepersberg, Crit. Rev. Biotechnol. 14;251-285 (1994). For a review of cloned clusters of antibiotic-producing genes, see Chater Bio/Technology 8:115-121 (1990). Other examples of antibiotic synthesis genes transferred to industrial producing strains, or over expression of genes, include tylosin, cephamycin C, cephalosporin C, LL-E33288 complex (an antitumor and antibacterial agent), doxorubicin, spiramycin and other macrolide antibiotics, reviewed in

Cameron et al. Appl. Biochem. Biotechnol. 38:105-140 (1993).

b. Biosynthesis to Replace Chemical Synthesis of Antibiotics

Some antibiotics are currently made by chemical modifications of biologically produced starting compounds. Complete biosynthesis of the desired molecules may currently be impractical because of the lack of an enzyme with the required enzymatic activity and substrate specificity. For example, 7-aminodeacetooxycephalosporanic acid (7-ADCA) is a precursor for semi-synthetically produced cephalosporins. 7-ADCA is made by a chemical ring expansion from penicillin V followed by enzymatic deacylation of the phenoxyacetal group. Cephalosporin V could in principle be produced biologically from penicillin V using penicillin N expandase, but penicillin V is not used as a substrate by any known expandase. The GEOS recombination techniques of the invention can be used to alter the enzyme so that it will use penicillin V as a substrate. Similarly, penicillin transacylase could be so modified to accept cephalosporins or cephamycins as substrates.

In yet another example, penicillin amidase expressed in E. coli is a key enzyme in the production of penicillin G derivatives. The enzyme is generated from a precursor peptide and tends to accumulate as insoluble aggregates in the periplasm unless non-metabolizable sugars are present in the medium (Scherrer et al. Appl. Microbiol. Biotechnol. 42:85-91 (1994)). Evolution of this enzyme through the methods of the instant invention could be used to generate an enzyme that folds better, leading to a higher level of active enzyme expression.

In yet another example, Penicillin G acylase covalently linked to agarose is used in the synthesis of penicillin G derivatives. The enzyme can be stabilized for increased activity, longevity and/or thermal stability by chemical modification (Femandez-Lafuente et. al. Enzyme Microb. Technol. 14:489-495 (1992). Increased thermal stability is an especially attractive application of the GEOS recombination techniques of the instant invention, which can obviate the need for the chemical modification of such enzymes. Selection for thermostability can be performed in vivo in E. coli or in thermophiles at higher temperatures. In general, thermostability is a good first step in enhancing general stabilization of enzymes.

c. Polyketides

Polyketides include antibiotics such as tetracycline and erythromycin, anti-cancer agents such as daunomycin, immunosuppressants such as FK506 and rapamycin and veterinary products such as monesin and avermectin. Polyketide synthases (PKS's) are multifunctional enzymes that control the chain length, choice of chain-building units and reductive cycle that generates the huge variation in naturally occurring polyketides. Polyketides are built up by sequential transfers of "extender units" (fatty acyl CoA groups) onto the appropriate starter unit (examples are acetate, coumarate, propionate and malonamide). The PKS's determine the number of condensation reactions and the type of extender groups added and may also fold and cyclize the polyketide precursor. PKS's reduce specific beta-keto groups and may dehydrate the resultant beta-hydroxyls to form

double bonds. Modifications of the nature or number of building blocks used, positions at which beta-keto groups are reduced, the extent of reduction and different positions of possible cyclizations, result in formation of different final products. Polyketide research is currently focused on modification and inhibitor studies, site directed mutagenesis and 3-D structure elucidation to lay the groundwork for rational changes in enzymes that will lead to new polyketide products.

Recently, McDaniel et al. (Science 262:1546-1550 (1995)) have developed a Streptomyces host-vector system for efficient construction and expression of recombinant PKSs. Hutchinson (Bio/Technology 12:375-308 (1994)) reviewed targeted mutation of specific biosynthetic genes and suggested that microbial isolates can be screened by DNA hybridization for genes associated with known pharmacologically active agents so as to provide new metabolites and large amounts of old ones. In particular, that review focuses on polyketide synthase and pathways to aminoglycoside and oligopeptide antibiotics.

The GEOS recombination techniques of the instant invention can be used to generate modified enzymes and enzyme clusters that produce novel polyketides without such detailed analytical effort. The availability of the PKS genes on plasmids and the existence of E. coli-Streptomyces shuttle vectors (Wehmeier Gene 165:149-150 (1995)) makes the process of GEOS recombination especially attractive by the techniques described herein. Techniques for selection of antibiotic producing organisms can be used as described herein; additionally, in some embodiments screening for a particular desired polyketide activity or compound is preferable.

d. Isoprenoids

Isoprenoids result from cyclization of farnesyl pyrophosphate by sesquiterpene synthases. The diversity of isoprenoids is generated not by the backbone, but by control of cyclization. Cloned examples of isoprenoid synthesis genes include trichodiene synthase from Fusarium sprorotrichioides, pentalene synthase from Streptomyces, aristolochene synthase from Penicillium roquefortii, and epi-aristolochene synthase from N. tabacum (Cane, D. E. (1995). Isoprenoid antibiotics, pages 633-655, in "Genetics and Biochemistry of Antibiotic Production" edited by Vining, L. C. & Stuttard, C., published by Butterworth-Heinemann). GEOS recombination of sesquiterpene synthases will be of use both in allowing expression of these enzymes in heterologous hosts (such as plants and industrial microbial strains) and in alteration of enzymes to change the cyclized product made. A large number of isoprenoids are active as antiviral, antibacterial, antifungal, herbicidal, insecticidal or cytostatic agents. Antibacterial and antifungal isoprenoids could thus be preferably screened for using the indicator cell type system described herein, with the producing cell competing with bacteria or fungi for nutrients. Antiviral isoprenoids could be screened for preferably by their ability to confer resistance to viral attack on the producing cell.

e. Bioactive Peptide Derivatives

Examples of bioactive non-ribosomally synthesized peptides include the antibiotics

cyclosporin, pepstatin, actinomycin, gramicidin, depsipeptides, vancomycin, etc. These peptide derivatives are synthesized by complex enzymes rather than ribosomes. Again, increasing the yield of such non-ribosomally synthesized peptide antibiotics has thus far been done by genetic identification of biosynthetic "bottlenecks" and over expression of specific enzymes (See, for example, p. 133-135 in "Genetics and Biochemistry of Antibiotic Production" edited by Vining, L. C. & Stuttard, C., published by Butterworth-Heinemann). GEOS recombination of the enzyme clusters can be used to improve the yields of existing bioactive non-ribosomally made peptides in both natural and heterologous hosts. Like polyketide synthases, peptide synthases are modular and multifunctional enzymes catalyzing condensation reactions between activated building blocks (in this case amino acids) followed by modifications of those building blocks (see Kleinkauf, H. and von Dohren, H. Eur. J. Biochem. 236:335-351 (1996)). Thus, as for polyketide synthases, GEOS recombination can also be used to alter peptide synthases; modifying the specificity of the amino acid recognized by each binding site on the enzyme and altering the activity or substrate specificities of sites that modify these amino acids to produce novel compounds with antibiotic activity.

Other peptide antibiotics are made ribosomally and then post-translationally modified. Examples of this type of antibiotics are lantibiotics (produced by gram positive bacteria such Staphylococcus, Streptomyces, Bacillus, and Actinoplanes) and microcins (produced by Enterobacteriaceae). Modifications of the original peptide include (in lantibiotics) dehydration of serine and threonine, condensation of dehydroamino acids with cysteine, or simple N- and C-terminal blocking (microcins). For ribosomally made antibiotics both the peptide-encoding sequence and the modifying enzymes may have their expression levels modified by GEOS recombination. Again, this will lead to both increased levels of antibiotic synthesis, and by modulation of the levels of the modifying enzymes (and the sequence of the ribosomally synthesized peptide itself) novel antibiotics.

Screening can be done as for other antibiotics as described herein, including competition with a sensitive (or even initially insensitive) microbial species. Use of competing bacteria that have resistances to the antibiotic being produced will select strongly either for greatly elevated levels of that antibiotic (so that it swamps out the resistance mechanism) or for novel derivatives of that antibiotic that are not neutralized by the resistance mechanism.

f. Polymers

Several examples of metabolic engineering to produce biopolymers have been reported, including the production of the biodegradable plastic polyhydroxybutarate (PHB), and the polysaccharide xanthan gum. For a review, see Cameron et al. Applied Biochem. Biotech. 38:105-140 (1993). Genes for these pathways have been cloned, making them excellent candidates for the GEOS recombination techniques described herein. Expression of such evolved genes in a commercially viable host such as E. coli is an especially attractive application of this technology.

Examples of starting materials for GEOS recombination include but are not limited to genes from bacteria such as Alcaligenes, Zoogloea, Rhizobium, Bacillus, and Azobacter,

which produce polyhydroxyalkanoates (PHAs) such as polyhyroxybutyrate (PHB) intracellularly as energy reserve materials in response to stress. Genes from Alcaligenes eutrophus that encode enzymes catalyzing the conversion of acetoacetyl CoA to PHB have been transferred both to E. coli and to the plant Arabidopsis thaliana (Poirier et al. Science 256:520-523 (1992)). Two of these genes (phbB and phbC, encoding acetoacetyl-CoA reductase and PHB synthase respectively) allow production of PHB in Arabidopsis. The plants producing the plastic are stunted, probably because of adverse interactions between the new metabolic pathway and the plants' original metabolism (i.e., depletion of substrate from the mevalonate pathway). Improved production of PHB in plants has been attempted by localization of the pathway enzymes to organelles such as plastids. Other strategies such as regulation of tissue specificity, expression timing and cellular localization have been suggested to solve the deleterious effects of PHB expression in plants. The GEOS recombination techniques of the invention can be used to modify such heterologous genes as well as specific cloned interacting pathways (e.g., mevalonate), and to optimize PHB synthesis in industrial microbial strains, for example to remove the requirement for stresses (such as nitrogen limitation) in growth conditions.

Additionally, other microbial polyesters are made by different bacteria in which additional monomers are incorporated into the polymer (Peoples et al. in Novel Biodegradable Microbial Polymers, E A Dawes, ed., pp191-202 (1990)). Application of the subject GEOS recombination method to these genes or pathways singly or in combination into a heterologous host will allow the production of a variety of polymers with differing properties, including variation of the monomer subunit ratios in the polymer. Another polymer whose synthesis may be manipulated by GEOS recombination is cellulose. The genes for cellulose biosynthesis have been cloned from Agrobacterium tumefaciens (Matthysse, A. G. et. al. J. Bacteriol. 177:1069-1075 (1995)). GEOS recombination of this biosynthetic pathway could be used either to increase synthesis of cellulose, or to produce mutants in which alternative sugars are incorporated into the polymer.

g. Carotenoids

Carotenoids are a family of over 600 terpenoids produced in the general isoprenoid biosynthetic pathway by bacteria, fungi and plants (for a review, see Armstrong, J. Bact. 176:4795-4802 (1994)). These pigments protect organisms against photooxidative damage as well as functioning as anti-tumor agents, free radical-scavenging anti-oxidants, and enhancers of the immune response. Additionally, they are used commercially in pigmentation of cultured fish and shellfish. Examples of carotenoids include but are not limited to myxobacton, spheroidene, spheroidenone, lutein, astaxanthin, violaxanthin, 4-ketorulene, myxoxanthrophyll, echinenone, lycopene, zeaxanthin and its mono- and di-glucosides, alpha-, beta-, gamma- and delta-carotene, beta-cryptoxanthin monoglucoside and neoxanthin.

Carotenoid synthesis is catalyzed by relatively small numbers of clustered genes: 11 different genes within 12 kb of DNA from Myxococcus xanthus (Botella et al. Eur. J. Biochem. 233:238-248 (1995)) and 8 genes within 9 kb of DNA from Rhodobacter sphaeroides (Lang et. al. J. Bact. 177:2064-2073 (1995)). In some microorganisms, such as

Thermus thermophilus, these genes are plasmid-borne (Tabata et al. FEBS Letts 341:251-255 (1994)). These features make carotenoid synthetic pathways especially attractive candidates for GEOS recombination.

Transfer of some carotenoid genes into heterologous organisms results in expression. For example, genes from Erwina uredovora and Haematococcus pluvialis will function together in E. coli (Kajiwara et al. Plant Mol. Biol. 29:343-352 (1995)). E. herbicola genes will function in R. sphaeroides (Hunter et al. J. Bact. 176:3692-3697 (1994)). However, some other genes do not; for example, R. capsulatus genes do not direct carotenoid synthesis in E. coli (Marrs, J. Bact. 146:1003-1012 (1981)).

In an embodiment of the invention, the GEOS recombination techniques of the invention can be used to generate variants in the regulatory and/or structural elements of genes in the carotenoid synthesis pathway, allowing increased expression in heterologous hosts. Indeed, traditional techniques have been used to increase carotenoid production by increasing expression of a rate limiting enzyme in Thermus thermophilus (Hoshino et al. Appl. Environ. Micro. 59:3150-3153 (1993)). Furthermore, mutation of regulatory genes can cause constitutive expression of carotenoid synthesis in actinomycetes, where carotenoid photoinducibility is otherwise unstable and lost at a relatively high frequency in some species (Kato et al. Mol. Gen. Genet. 247:387-390 (1995)). These are both mutations that can be obtained by GEOS recombination.

The GEOS recombination techniques of the invention as described herein can be used to evolve one or more carotenoid synthesis genes in a desired host without the need for analysis of regulatory mechanisms. Since carotenoids are colored, a calorimetric assay in microtiter plates, or even on growth media plates, can be used for screening for increased production.

In addition to increasing expression of carotenoids, carotenogenic biosynthetic pathways have the potential to produce a wide diversity of carotenoids, as the enzymes involved appear to be specific for the type of reaction they will catalyze, but not for the substrate that they modify. For example, two enzymes from the marine bacterium Agrobacterium aurantiacum (CrtW and CrtZ) synthesize six different ketocarotenoids from beta-carotene (Misawa et al. J. Bact. 177:6576-6584 (1995)). This relaxed substrate specificity means that a diversity of substrates can be transformed into an even greater diversity of products. Introduction of foreign carotenoid genes into a cell can lead to novel and functional carotenoid-protein complexes, for example in photosynthetic complexes (Hunter et al. J.Bact. 176:3692-3697 (1994)). Thus, the deliberate recombination of enzymes through the GEOS recombination techniques of the invention is likely to generate novel compounds. Screening for such compounds can be accomplished, for example, by the cell competition/survival techniques discussed above and by a calorimetric assay for pigmented compounds.

Another method of identifying new compounds is to use standard analytical techniques such as mass spectroscopy, nuclear magnetic resonance, high performance liquid chromatography, etc. Recombinant microorganisms can be pooled and extracts or media

supernatants assayed from these pools. Any positive pool can then be subdivided and the procedure repeated until the single positive is identified ("sib-selection").

h. Indigo Biosynthesis

Many dyes, i.e. agents for imparting color, are specialty chemicals with significant markets. As an example, indigo is currently produced chemically. However, nine genes have been combined in E. coli to allow the synthesis of indigo from glucose via the tryptophan/indole pathway (Murdock et al. Bio/Technology 11:381-386 (1993)). A number of manipulations were performed to optimize indigo synthesis: cloning of nine genes, modification of the fermentation medium and directed changes in two operons to increase reaction rates and catalytic activities of several enzymes. Nevertheless, bacterially produced indigo is not currently an economic proposition. The GEOS recombination techniques of the instant invention could be used to optimize indigo synthesizing enzyme expression levels and catalytic activities, leading to increased indigo production, thereby making the process commercially viable and reducing the environmental impact of indigo manufacture. Screening for increased indigo production can be done by calorimetric assays of cultures in microtiter plates.

i. Amino Acids

Amino acids of particular commercial importance include but are not limited to phenylalanine, monosodium glutamate, glycine, lysine, threonine, tryptophan and methionine. Backman et al. (Ann. NY Acad. Sci. 589:16-24 (1990)) disclosed the enhanced production of phenylalanine in E. coli via a systematic and downstream strategy covering organism selection, optimization of biosynthetic capacity, and development of fermentation and recovery processes.

As described in Simpson et al. (Biochem Soc Trans, 23;381-387 (1995)), current work in the field of amino acid production is focused on understanding the regulation of these pathways in great molecular detail. The GEOS recombination techniques of the instant invention would obviate the need for this analysis to obtain bacterial strains with higher secreted amino acid yields. Amino acid production could be optimized for expression using GEOS recombination of the amino acid synthesis and secretion genes as well as enzymes at the regulatory phosphoenolpyruvate branchpoint, from such organisms as Serratia marcescens, Bacillus, and the Corynebacterium-Brevibacterium group. In some embodiments of the invention, screening for enhanced production is preferably done in microtiter wells, using chemical tests well known in the art that are specific for the desired amino acid. Screening/selection for amino acid synthesis can also be done by using auxotrophic reporter cells that are themselves unable to synthesize the amino acid in question. If these reporter cells also produce a compound that stimulates the growth of the amino acid producer (this could be a growth factor, or even a different amino acid), then library cells that produce more amino acid will in turn receive more growth stimulant and will therefore grow more rapidly.

j. Vitamin C synthesis

L-Ascorbic acid (vitamin C) is a commercially important vitamin with a world production of over 35,000 tons in 1984. Most vitamin C is currently manufactured chemically by the Reichstein process, although recently bacteria have been engineered that are able to transform glucose to 2,5-keto-gluconic acid, and that product to 2-keto-L-idonic acid, the precursor to L-ascorbic acid (Boudrant, Enzyme Microb. Technol. 12:322-329 (1990)).

The efficiencies of these enzymatic steps in bacteria are currently low. Using the GEOS recombination techniques of the instant invention, the genes can be genetically engineered to create one or more operons followed by expression optimization of such a hybrid L-ascorbic acid synthetic pathway to result in commercially viable microbial vitamin C biosynthesis. In some embodiments, screening for enhanced L-ascorbic acid production is preferably done in microtiter plates, using assays well known in the art.

vi) Recombination of Genes For Bioremediation

Modern industry generates many pollutants for which the environment can no longer be considered an infinite sink. Naturally occurring microorganisms are able to metabolize thousands of organic compounds, including many not found in nature (e.g xenobiotics). Bioremediation, the deliberate use of microorganisms for the biodegradation of man-made wastes, is an emerging technology that offers cost and practicality advantages over traditional methods of disposal. The success of bioremediation depends on the availability of organisms that are able to detoxify or mineralize pollutants. Microorganisms capable of degrading specific pollutants can be generated by genetic engineering and GEOS recombination.

Although bioremediation is an aspect of pollution control, a more useful approach in the long term is one of prevention before industrial waste is pumped into the environment. Exposure of industrial waste streams to GEOS-generated microorganisms capable of degrading the pollutants they contain would result in detoxification of mineralization of these pollutants before the waste stream enters the environment. Issues of releasing recombinant organisms can be avoided by containing them within bioreactors fitted to the industrial effluent pipes. This approach would also allow the microbial mixture used to be adjusted to best degrade the particular wastes being produced. Finally, this method would avoid the problems of adapting to the outside world and dealing with competition that face many laboratory microorganisms.

In the wild, microorganisms have evolved new catabolic activities enabling them to exploit pollutants as nutrient sources for which there is no competition. However, pollutants that are present at low concentrations in the environment may not provide a sufficient advantage to stimulate the evolution of catabolic enzymes. For a review of such naturally occurring evolution of biodegradative pathways and the manipulation of some of microorganisms by classical techniques, see Ramos et al., Bio/Technology 12:1349-1355 (1994).

Generation of new catabolic enzymes or pathways for bioremediation has thus relied upon

deliberate transfer of specific genes between organisms (Wackett et al., supra), forced matings between bacteria with specific catabolic capabilities (Brenner et al. Biodegradation 5:359-377 (1994)), or prolonged selection in a chemostat. Some researchers have attempted to facilitate evolution via naturally occurring genetic mechanisms in their chemostat selections by including microorganisms with a variety of catabolic pathways (Kellogg et. al. Science 214:1133-1135 (1981); Chakrabarty American Society of Micro. Biol. News 62:130-137 (1996)). For a review of efforts in this area, see Cameron et al. Applied Biochem. Biotech. 38:105-140 (1993).

Current efforts in improving organisms for bioremediation take a labor-intensive approach in which many parameters are optimized independently, including transcription efficiency from native and heterologous promoters, regulatory circuits and translational efficiency as well as improvement of protein stability and activity (Timmis et al. Ann. Rev. Microbiol. 48:525-527 (1994)).

A GEOS recombination approach overcomes a number of limitations in the bioremediation capabilities of naturally occurring microorganisms. Both enzyme activity and specificity can be altered, simultaneously or sequentially, by the methods of the invention. For example, novel catabolic enzymes can be created to increase the rate at which they act on a substrate. Although knowledge of a rate-limiting step in a metabolic pathway is not required to practice the invention, rate-limiting proteins in pathways can be evolved to have increased expression and/or activity, the requirement for inducing substances can be eliminated, and enzymes can be evolved that catalyze novel reactions.

Some examples of chemical targets for bioremediation include but are not limited to benzene, xylene, and toluene, camphor, naphthalene, halogenated hydrocarbons, polychlorinated biphenyls (PCBs), trichlorethylene, pesticides such as pentachlorophenyls (PCPs), and herbicides such as atrazine.

a) Aromatic Hydrocarbons

Preferably, when an enzyme is "evolved" to have a new catalytic function, that function is expressed, either constitutively or in response to the new substrate. The target recombination method subjects both structural and regulatory elements (including the structure of regulatory proteins) of a protein to recombinogenic mutagenesis simultaneously. Selection of mutants that are efficiently able to use the new substrate as a nutrient source will be sufficient to ensure that both the enzyme and its regulation are optimized, without detailed analysis of either protein structure or operon regulation.

Examples of aromatic hydrocarbons include but are not limited to benzene, xylene, toluene, biphenyl, and polycyclic aromatic hydrocarbons such as pyrene and naphthalene. These compounds are metabolized via catechol intermediates. Degradation of catechol by Pseudomonas putida requires induction of the catabolic operon by cis, cis-muconate which acts on the CatR regulatory protein. The binding site for the CatR protein is G-N11-A, while the optimal sequence for the LysR class of activators (of which CatR is a member) is T-N11-A. Mutation of the G to a T in the CatR binding site enhances the expression of

catechol metabolizing genes (Chakrabarty, American Society of Microbiology News 62:130-137 (1996)). This demonstrates that the control of existing catabolic pathways is not optimized for the metabolism of specific xenobiotics, and suggests that the subject method can be used to generate recombinant bacteria that are better able to degrade the target compound.

As an example of starting materials, dioxygenases are required for many pathways in which aromatic compounds are catabolized. Even small differences in dioxygenase sequence can lead to significant differences in substrate specificity (Furukawa et al. J. Bact. 175:5224-5232 (1993); Erickson et al. App. Environ. Micro. 59:3858-3862 (1993)). A hybrid enzyme made using sequences derived from two or more "parental" enzymes may possess catalytic activities that are intermediate between the parents (Erickson, ibid.), or may actually be better than either parent for a specific reaction (Furukawa et al. J. Bact. 176:2121-2123 (1994)). For example, a four subunit enzyme can be produced by expressing two or more subunits from different dioxygenases. Thus, sequences from one or more genes encoding dioxygenases can be used in the recombination techniques of the instant invention, to generate enzymes with new specificities. In addition, other features of the catabolic pathway can also be evolved using these techniques, simultaneously or sequentially, to optimize the metabolic pathway for an activity of interest.

b) Halogenated Hydrocarbons

Large quantities of halogenated hydrocarbons are produced annually for uses as solvents and biocides. These include, in the United States alone, over 5 million tons of both 1,2-dichloroethane and vinyl chloride used in PVC production in the U.S. alone. The compounds are largely not biodegradable by processes in single organisms, although in principle haloaromatic catabolic pathways can be constructed by combining genes from different microorganisms. Enzymes can be manipulated to change their substrate specificities. The subject method offers the possibility of tailoring enzyme specificity to new substrates without needing detailed structural analysis of the enzymes.

As an example of possible starting materials for the methods of the instant invention, Wackett et al. (Nature 368:627-629 (1994)) recently demonstrated that through classical techniques a recombinant Pseudomonas strain in which seven genes encoding two multi-component oxygenases are combined, generated a single host that can metabolize polyhalogenated compounds by sequential reductive and oxidative techniques to yield non-toxic products. These and/or related materials can be subjected to the combinatorial techniques discussed above so as to evolve and optimize a biodegradative pathway in a single organism.

Trichloroethylene is a significant groundwater contaminant. It is degraded by microorganisms in a cometabolic way (i.e., no energy or nutrients are derived). The enzyme must be induced by a different compound (e.g., Pseudomonas cepacia uses toluene-4-monoxygenase, which requires induction by toluene, to destroy trichloroethylene). Furthermore, the degradation pathway involves formation of highly reactive epoxides that can inactivate the enzyme (Timmis et al. Ann. Rev. Microbiol. 48:525-557 (1994)). The

GEOS recombination techniques of the invention could be used to generate libraries of genes having mutations to coding sequence enzymes and its regulatory region such that it is produced constitutively, and is less susceptible to epoxide inactivation. In some embodiments of the invention, selection of hosts constitutively producing the enzyme and less susceptible to the epoxides can be accomplished by demanding growth in the presence of increasing concentrations of trichloroethylene in the absence of inducing substances.

c) Polychlorinated Biphenyls (PCBs) and Polycyclic Aromatic Hydrocarbons (PAHs)

PCBs and PAHs are families of structurally related compounds that are major pollutants at many Superfund sites. Bacteria transformed with plasmids encoding enzymes with broader substrate specificity have been used commercially. In nature, no known pathways have been generated in a single host that degrade the larger PAHs or more heavily chlorinated PCBs. Indeed, often the collaboration of anaerobic and aerobic bacteria are required for complete metabolism.

Thus, likely sources for starting material for GEOS recombination include identified genes encoding PAH-degrading catabolic pathways on large (20-100 KB) plasmids (Sanseverino et al. Applied Environ. Micro. 59;1931-1937 (1993); Simon et al. Gene 127;31-37 (1993); Zylstra et al. Annals of the NY Acad. Sci. 721;386-398 (1994)); while biphenyl and PCB-metabolizing enzymes are encoded by chromosomal gene clusters, and in a number of cases have been cloned onto plasmids (Hayase et al. J. Bacteriol. 172;1160-1164 (1990); Furukawa et al. Gene 98;21-28 (1992); Hofer et al. Gene 144;9-16 (1994)). The materials can be subjected to the techniques discussed above so as to evolve a biodegradative pathway in a single organism.

Substrate specificity in the PCB pathway largely results from enzymes involved in initial dioxygenation reactions, and can be significantly altered by mutations in those enzymes (Erickson et al. Applied Environ. Micro. 59;3858-38662 (1993); Furukawa et al. J. Bact. 175;5224-5232 (1993). Mineralization of PAHs and PCBs requires that the downstream pathway is able to metabolize the products of the initial reaction (Brenner et al. Biodegradation 5;359-377 (1994)). In this case, application of the subject method to the entire pathway with selection for bacteria able to use the PCB or PAH as the sole carbon source can allow production of novel PCB and PAH degrading bacteria.

d) Herbicides

A general method for evolving genes for the catabolism of insoluble herbicides is exemplified as follows for atrazine. Atrazine [2-chloro-4-(ethylamino)-6-(isopropylamino)-1,3,5-triazine] is a moderately persistent herbicide which is frequently detected in ground and surface water at concentrations exceeding the 3 ppb health advisory level set by the EPA. Atrazine can be slowly metabolized by a Pseudomonas species (Mandelbaum et al. Appl. Environ. Micro. 61:1451-1457 (1995)). The enzymes catalyzing the first two steps in atrazine metabolism by Pseudomonas are encoded by genes AtzA and AtzB (de Souza et al. Appl. Environ. Micro. 61;3373-3378 (1995)). These genes can be cloned from various species. E. coli engineered with these genes convert atrazine to much more soluble

metabolites. It is thus possible to screen for enzyme activity by growing bacteria on plates containing atrazine. The herbicide forms an opaque precipitate in the plates, but cells expressing the AtzA and AtzB genes secrete atrazine degrading enzymes, leading to a clear halo around those cells or colonies. Typically, the size of the halo and the rate of its formation can be used to assess the level of activity so that picking colonies with the largest halos allows selection of the more active or highly produced atrazine degrading enzymes. Thus, this pathway can be subjected to GEOS sequence recombination formats described above to optimize the catabolism of atrazine in E. coli or another host of choice, including Pseudomonas. Screening of host colonies expressing the evolved genes can be done on agar plates containing atrazine to observe halo formation. This is a generally applicable method for screening enzymes that metabolize insoluble compounds to those that are soluble (e.g., polycyclicaromatic hydrocarbons). Additionally, catabolism of atrazine can provide a source of nitrogen for the cell; if no other nitrogen is available, cell growth will be limited by the rate at which the cells can catabolize nitrogen. Cells able to utilize atrazine as a nitrogen source can thus be selected from a background of non-utilizers or poor-utilizers.

e) Heavy Metal Detoxification

Bacteria are used commercially to detoxify arsenate waste generated by the mining of arsenopyrite gold ores. As well as mining effluent, industrial waste water is often contaminated with heavy metals (e.g., those used in the manufacture of electronic components and plastics). Thus, simply to be able to perform other bioremedial functions, microorganisms must be resistant to the levels of heavy metals present, including mercury, arsenate, chromate, cadmium, silver, etc.

A strong selective pressure is the ability to metabolize a toxic compound to one less toxic. Heavy metals are toxic largely by virtue of their ability to denature proteins (Ford et al, Bioextraction and Biodeterioration of Metals, p. 1-23). Detoxification of heavy metal contamination can be effected in a number of ways including changing the solubility or bioavailability of the metal, changing its redox state (e.g. toxic mercuric chloride is detoxified by reduction to the much more volatile elemental mercury) and even by bioaccumulation of the metal by immobilized bacteria or plants. The accumulation of metals to a sufficiently high concentration allows metal to be recycled; smelting burns off the organic part of the organism, leaving behind reusable accumulated metal. Resistances to a number of heavy metals (arsenate, cadmium, cobalt, chromium, copper, mercury, nickel, lead, silver, and zinc) are plasmid encoded in a number of species including Staphylococcus and Pseudomonas (Silver et al, Environ. Health Perspect. 102:107-113 (1994); Ji et al, J. Ind. Micro. 14:61-75 (1995)). These genes also confer heavy metal resistance on other species as well (e.g., E. coli). The GEOS recombination techniques of the instant invention can be used to increase microbial heavy metal tolerances, as well as to increase the extent to which cells will accumulate heavy metals. For example, the ability of E. coli to detoxify arsenate can be improved.

Cyanide is very efficiently used to extract gold from rock containing as little as 0.2 oz per ton. This cyanide can be microbially neutralized and used as a nitrogen source by fungi or bacteria such as Pseudomonas fluorescens. A problem with microbial cyanide degradation

is the presence of toxic heavy metals in the leachate. GEOS can be used to increase the resistance of bioremedial microorganisms to toxic heavy metals, so that they will be able to survive the levels present in many industrial and Superfund sites. This will allow them to biodegrade organic pollutants including but not limited to aromatic hydrocarbons, halogenated hydrocarbons, and biocides.

f) Microbial Mining

"Bioleaching" is the process by which microbes convert insoluble metal deposits (usually metal sulfides or oxides) into soluble metal sulfates. Bioleaching is commercially important in the mining of arsenopyrite, but has additional potential in the detoxification and recovery of metals and acids from waste dumps. Naturally occurring bacteria capable of bioleaching are reviewed by Rawlings and Silver (Bio/Technology 13:773-778 (1995)). These bacteria are typically divided into groups by their preferred temperatures for growth. The more important mesophiles are Thiobacillus and Leptospirillum species. Moderate thermophiles include Sulfobacillus species. Extreme thermophiles include Sulfolobus species. Many of these organisms are difficult to grow in commercial industrial settings, making their catabolic abilities attractive candidates for transfer to and optimization in other organisms such as Pseudomonas, Rhodococcus, T. ferrooxidans or E. coli. Genetic systems are available for at least one strain of T. ferrooxidans, allowing the manipulation of its genetic material on plasmids.

The GEOS recombination methods described above can be used to optimize the catalytic abilities in native hosts or heterologous hosts for evolved bioleaching genes or pathways, such as the ability to convert metals from insoluble to soluble salts. In addition, leach rates of particular ores can be improved as a result of, for example, increased resistance to toxic compounds in the ore concentrate, increased specificity for certain substrates, ability to use different substrates as nutrient sources, and so on.

g) Oil Desulfurization

The presence of sulfur in fossil fuels has been correlated with corrosion of pipelines, pumping, and refining equipment, and with the premature breakdown of combustion engines. Sulfur also poisons many catalysts used in the refining of fossil fuels. The atmospheric emission of sulfur combustion products is known as acid rain.

Microbial desulfurization is an appealing bioremediation application. Several bacteria have been reported that are capable of catabolizing dibenzothiophene (DBT), which is the representative compound of the class of sulfur compounds found in fossil fuels. U.S. Pat. No. 5,356,801, for example, discloses the cloning of a DNA molecule from Rhodococcus rhodochrous capable of biocatalyzing the desulfurization of oil. Denome et al. (Gene 175:6890-6901 (1995)) disclose the cloning of a 9.8 kb DNA fragment from Pseudomonas encoding the upper naphthalene catabolizing pathway which also degrades dibenzothiophene. Other genes have been identified that perform similar functions (such as disclosed in U.S. Pat. No. 5,356,801).

The activity of these enzymes is currently too low to be commercially viable, but the pathway could be increased in efficiency using the GEOS recombination techniques of the invention. The desired property of the genes of interest is their ability to desulfurize dibenzothiophene. In some embodiments of the invention, selection is preferably accomplished by coupling this pathway to one providing a nutrient to the bacteria. Thus, for example, desulfurization of dibenzothiophene results in formation of hydroxybiphenyl. This is a substrate for the biphenyl-catabolizing pathway which provides carbon and energy. Selection would thus be done by "shuffling" the dibenzothiophene genes and transforming them into a host containing the biphenyl-catabolizing pathway. Increased dibenzothiophene desulfurization will result in increased nutrient availability and increased growth rate.

h) Organo-Nitro Compounds

Organo-nitro compounds are used as explosives, dyes, drugs, polymers and antimicrobial agents. Biodegradation of these compounds occurs usually by way of reduction of the nitrate group, catalyzed by nitroreductases, a family of broadly-specific enzymes. Partial reduction of organo-nitro compounds often results in the formation of a compound more toxic than the original (Hassan et al. 1979 Arch Bioch Biop. 196;385-395). GEOS recombination of nitroreductases can produce enzymes that are more specific, and able to more completely reduce (and thus detoxify) their target compounds (examples of which include but are not limited to nitrotoluenes and nitrobenzenes). Nitro-reductases can be isolated from bacteria isolated from explosive-contaminated soils, such as Morganella morganii and Enterobacter cloacae (Bryant et. al., 1991, J. Biol Chem. 266:4126-4130). A preferred selection method is to look for increased resistance to the organo-nitro compound of interest, since that will indicate that the enzyme is also able to reduce any toxic partial reduction products of the original compound.

vii) Use of Alternative Substrates for Chemical Synthesis

Metabolic engineering can be used to alter microorganisms that produce industrially useful chemicals, so that they will grow using alternate and more abundant sources of nutrients, including human-produced industrial wastes. This typically involves providing both a transport system to get the alternative substrate into the engineered cells and catabolic enzymes from the natural host organisms to the engineered cells. In some instances, enzymes can be secreted into the medium by engineered cells to degrade the alternate substrate into a form that can more readily be taken up by the engineered cells; in other instances, a batch of engineered cells can be grown on one preferred substrate, then lysed to liberate hydrolytic enzymes for the alternate substrate into the medium, while a second inoculum of the same engineered host or a second host is added to utilize the hydrolyzate.

The starting materials for the subject recombination method will typically be genes for utilization of a substrate or its transport. Examples of nutrient sources of interest include but are not limited to lactose, whey, galactose, mannitol, xylan, cellobiose, cellulose and sucrose, thus allowing cheaper production of compounds including but not limited to ethanol, tryptophan, rhamnolipid surfactants, xanthan gum, and polyhydroxylalkanoate.

For a review of such substrates as desired target substances, see Cameron et al. (Appl. Biochem. Biotechnol. 38:105-140 (1993)).

The GEOS recombination methods described herein can be used to optimize the ability of native hosts or heterologous hosts to utilize a substrate of interest, to evolve more efficient transport systems, to increase or alter specificity for certain substrates, and so on.

viii) Modification of Cell Properties.

Although not strictly examples of manipulation of intermediary metabolism, GEOS recombination techniques can be used to improve or alter other aspects of cell properties, from growth rate to ability to secrete certain desired compounds to ability to tolerate increased temperature or other environmental stresses. Some examples of traits engineered by traditional methods include expression of heterologous proteins in bacteria, yeast, and other eukaryotic cells, antibiotic resistance, and phage resistance. Any of these traits is advantageously evolved by the GEOS recombination techniques of the instant invention. Examples include replacement of one nutrient uptake system (e.g. ammonia in Methylophilus methylotrophus) with another that is more energy efficient; expression of haemoglobin to improve growth under conditions of limiting oxygen; redirection of toxic metabolic end products to less toxic compounds; expression of genes conferring tolerance to salt, drought and toxic compounds and resistance to pathogens, antibiotics and bacteriophage, reviewed in Cameron et. al. Appl Biochem Biotechnol, 38:105-140 (1993).

The heterologous genes encoding these functions all have the potential for further optimization in their new hosts by existing GEOS recombination technology. Since these functions increase cell growth rates under the desired growth conditions, optimization of the genes by "evolution" can simply involve "shuffling" the DNA and selecting the recombinants that grow faster with limiting oxygen, higher toxic compound concentration or whatever restrictive condition is being overcome.

Cultured mammalian cells also require essential amino acids to be present in the growth medium. This requirement could also be circumvented by expression of heterologous metabolic pathways that synthesize these amino acids (Rees et al. Biotechnology 8:629-633 (1990). GEOS recombination would provide a mechanism for optimizing the expression of these genes in mammalian cells. Once again, a preferred selection would be for cells that can grow in the absence of added amino acids.

Yet another candidate for improvement through the techniques of the invention is symbiotic nitrogen fixation. Genes involved in nodulation (nod, ndv), nitrogen reduction (nif, fix), host range determination (nod, hsp), bacteriocin production (tfx), surface polysaccharide synthesis (exo) and energy utilization (dct, hup) which have been identified (Paau, Biotech. Adv. 9:173-184 (1991)).

The main function of GEOS recombination in this case is in improving the survival of strains that are already known to be better nitrogen fixers. These strains tend to be less good at competing with strains already present in the environment, even though they are

better at nitrogen fixation. Targets for GEOS recombination such as nodulation and host range determination genes can be modified and selected for by their ability to grow on the new host. Similarly any bacteriocin or energy utilization genes that will improve the competitiveness of the strain will also result in greater growth rates. Selection can simply be performed by subjecting the target genes to GEOS recombination and forcing the inoculant to compete with wild type nitrogen fixing bacteria. The better the nitrogen fixing bacteria grow in the new host, the more copies of their recombined genes will be present for the next round of recombination. This growth rate differentiating selection is described herein in detail.

ix) Biodetectors/Biosensors

Bioluminescence or fluorescence genes can be used as reporters by fusing them to specific regulatory genes (Cameron et. al. Appl Biochem Biotechnol. 38:105-140 (1993)). A specific example is one in which the luciferase genes luxCDABE of Vibrio fischeri were fused to the regulatory region of the isopropylbenzene catabolism operon from Pseudomonas putida RE204. Transformation of this fusion construct into E. coli resulted in a strain which produced light in response to a variety of hydrophobic compound such as substituted benzenes, chlorinated solvents and naphthalene (Selifonova et. al., Appl Environ Microbiol 62:778-783 (1996)). This type of construct is useful for the detection of pollutant levels, and has the added benefit of only measuring those pollutants that are bioavailable (and therefore potentially toxic). Other signal molecules such as jellyfish green fluorescent protein could also be fused to genetic regulatory regions that respond to chemicals in the environment. This should allow a variety of molecules to be detected by their ability to induce expression of a protein or proteins which result in light, fluorescence or some other easily detected signal.

GEOS recombination can be used in several ways to modify this type of biodetection system. It can be used to increase the amplitude of the response, for example by increasing the fluorescence of the green fluorescent protein. GEOS recombination could also be used to increase induced expression levels or catalytic activities of other signal-generating systems, for example of the luciferase genes.

GEOS recombination can also be used to alter the specificity of biosensors. The regulatory region, and transcriptional activators that interact with this region and with the chemicals that induce transcription can also be shuffled. This should generate regulatory systems in which transcription is activated by analogues of the normal inducer, so that biodetectors for different chemicals can be developed. In this case, selection would be for constructs that are activated by the (new) specific chemical to be detected. Screening could be done simply with fluorescence (or light) activated cell sorting, since the desired improvement is in light production.

In addition to detection of environmental pollutants, biosensors can be developed that will respond to any chemical for which there are receptors, or for which receptors can be evolved by GEOS recombination, such as hormones, growth factors, metals and drugs. These receptors may be intracellular and direct activators of transcription, or they may be

membrane bound receptors that activate transcription of the signal indirectly, for example by a phosphorylation cascade. They may also not act on transcription at all, but may produce a signal by some post-transcriptional modification of a component of the signal generating pathway. These receptors may also be generated by fusing domains responsible for binding different ligands with different signaling domains. Again, GEOS recombination can be used to increase the amplitude of the signal generated to optimize expression and functioning of chimeric receptors, and to alter the specificity of the chemicals detected by the receptor.

IV Examples

The following examples are by way of illustration and are not intended to limit the claims. Persons of skill will readily recognize that the protocols of the examples can be modified in numerous non-critical ways.

Simultaneous assembly of a viable plasmid vector

To demonstrate the simultaneous assembly of multiple nucleic acid components having unique, non-palindromic terminal sequences, to produce a viable plasmid vector, three nucleic acid components are used. The first nucleic acid component is a gene coding for green fluorescent protein, 0.7 Kb in length, the second one is a 0.6 Kb molecule coding for terminator sequences and a histidine tag, and the third one is a 2.5 Kb molecule coding for the lac promoter, an ampicillin resistance gene, and an origin of replication.

1. Synthesis of the Nucleic Acid Components

The nucleic acid components used in the present example are synthesized by PCR amplification. The PCR reactions are performed in varying volumes (in general, 10-100 microliters) containing a 50 mM KCl, 10 mM Tris-HCl (pH 8.4), 1.5 mM MgCl.sub.2 buffer and 0.2 mM of each dNTP, 1.25 units of taq DNA polymerase, 10.sup.-5 M template molecules, and 20 pmol of each primer. The primers used contain uracil residues at specific locations in order to generate 3' terminal sequences as described in U.S. Pat. No. 5,137,814. The PCR reaction is carried out using a thermal cycling instrument, where there is an initial denaturation phase of 95.degree. C. for 5 minutes, followed by multiple cycles (20-40 cycles) of a denaturation step at 94.degree. C., an annealing step at 37-65.degree. C. and an extension step at 72.degree. C. The resulting PCR products are analyzed by gel electrophoresis to determine size and purity.

2. Generation of Terminal sequences

Following PCR amplification and purification of the correct size fragments, the PCR products (approximately 100-200 ng) are dissolved in 10 microliters of the UDG reaction buffer (25 mM Tris-HCl (pH 7.8), 10 mM Mg.sub.2 Cl, 4 mM beta-mercaptoethanol, 0.4 mM ATP). Single-stranded 3' Terminal sequences are made by treatment of the PCR product with 1-2 units of uracil DNA glycosidase (UDG) for 10 minutes at 37.degree. C. The enzyme is inactivated and reaction is terminated by heating the sample at 65.degree. C.

for 10 minutes.

3. Assembly and Ligation of the Nucleic Acid Components

To assemble the vector the individual purified nucleic acid components are mixed in equimolar amounts (approximately 20-200 ng total in 20 microliters) in the UDG treatment buffer and heated to 65.degree. C., followed by gradually cooling down to room temperature (25.degree. C.), to permit efficient annealing of the complementary ends of the nucleic acid components. The reaction mixture may optionally be treated with T4 DNA ligase at 14.degree. C. overnight to ligate the nucleic acid components or used directly to transform competent bacterial hosts.

4. Transformation

A 10 .mu.l aliquot of the assembled vector is added to 100 .mu.l of competent E. coli cells (DH5.alpha.), transformed following the manufacturers recommendations, and plated on LB plates containing ampicillin and IPTG.

5. Analysis of the Vector Construct

Isolated fluorescent colonies are selected and pure DNA plasmid prepared using a mini-prep. Correct assembly of the vector construct is determined using standard molecular biological methods, such as restriction enzyme digestion and agarose gel electrophoresis.

All of the above-cited references and publications are hereby incorporated by reference.

V. Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

Document comparison done by DeltaView on Friday, November 04, 2005 4:08:02 PM

Input:	
Document 1	iManageDeskSite://exchdms/Exchange/4007659/1
Document 2	iManageDeskSite://exchdms/Exchange/4007660/1
Rendering set	CHS Standard B&W

Legend:	
Insertion	
~~Deletion~~	
~~Moved from~~	
Moved to	
Style change	
Format change	
~~Moved deletion~~	
Inserted cell	
Deleted cell	
Moved cell	
Split/Merged cell	
Padding cell	

Statistics:	
	Count
Insertions	245
Deletions	129
Moved from	0
Moved to	0
Style change	0
Format changed	0
Total changes	374

Menu

Newsletter Home

2009 Newsletter

2008 Newsletter

2007 Newsletter

2006 Newsletter

2005 Newsletter

2004 Newsletter

2003 Newsletter

2002 Newsletter

2001 Newsletter

2000 Newsletter

1999 Newsletter

1998 Newsletter

New Faculty

Visiting Lecturer

Dr. Wynston: Retirement

Nobel Laureate: Cech

Distinguished Alumna

Remarks by the Chair

1998 In Memoriam

In Memoriam

Larry Copeland, BS Chemistry 1967, MS Chemistry 1970 (with Dr. A. G. Tharp), passed away on October 23, 1997. He was a long-term employee of Pilot Chemical Company in Santa Fe Springs, Calif., a manufacturer of detergents. He subsequently joined Rykoff-Sexton, Inc., Los Angeles, a major restaurant supplier, as Technical Director of the Detergents Division. Larry was an active and valued member of the Chemistry & Biochemistry Advisory Council of our department at the time of his death. He is survived by Jan Schrick Copeland, BS Chemistry 1968, a chemist with U.S. Borax, and two children, Erin and Sean, all of Long Beach. Both children are currently students at CSULB; Erin is a junior Psychology major and Sean is a freshman majoring in Theater Arts.

Dr. Peter D. Harney passed away on September 20, 1997. **Peter** received his BA in Chemistry at CSULB in 1975, his MS in Biochemistry and Biophysics from the University of Hawaii in 1983, and his PhD in Microbiology from USC in 1995. He was employed by Biogen Research Corporation in Cambridge, Mass. and subsequently with Diagnostic Products Corp., in Los Angeles. From 1986-92 he was a reseacher in the Norris Cancer Center at USC. In 1992 he joined the Immunotherapy Division & Gene Therapy Business

Unit of Baxter Healthcare Corp. in Irvine, Calif. as a Research Scientist. **Peter** also carried on an active research association with Dr. Roger Acey of our department which continued until his death. He is survived by his wife, Jennifer **Harney**; his children, Matthew, Christina and Brian **Harney**, all of **Aliso Viejo**, Calif.

Raymond Lyons, MD passed away on July 10, 1997 in St. Louis, Mo. as a result of an accident. A 1978 BA graduate of the Department of Chemistry and Biochemistry, he attended medical school at the University of Minnesota, and then practiced medicine in Minneapolis. In 1986 he moved to St. Louis in order to pursue his medical residency at St. Louis University Hospital in neurology and nuclear medicine. Complications due to multiple sclerosis cut short his medical career, and at the time of his death he resided at the Lutheran Convalescent Home in St. Louis. He is survived by his wife, Eugenie and two daughters, Elise and Rae, all of Minneapolis; and a son, Mitchell, of Lakewood, Calif.

Jerryl W. Neher. Word has been received from his sister that Jerryl W. Neher, BS Chemistry 1976, passed away on October 10, 1997.

Updated on 06/27/2006 11:33 AM by webmaster

1991 Newsletter (pdf)

1990 Newsletter (pdf)

1989 Newsletter (pdf)

1988 Newsletter (pdf)

1987 Newsletter (pdf)

1986 Newsletter (pdf)

1985 Newsletter (pdf)

1984 Newsletter (pdf)

1983 Newsletter (pdf)

1982 Newsletter (pdf)

1981 Newsletter (pdf)

1980 Newsletter (pdf)

1979 Newsletter (pdf)

1978 Newsletter (pdf)

1977 Newsletter (pdf)

1976 Newsletter (pdf)

Patent Application Information Retrieval

Order Certified Application As Filed | Order Certified File Wrapper | View Order List

08/488,015 INTRON-MEDIATED RECOMBINANT TECHNIQUES AND REAGENTS

Transaction History

Date	Transaction Description
03-16-2006	Correspondence Address Change
03-04-2006	Correspondence Address Change
04-29-1999	Mail Miscellaneous Communication to Applicant
04-29-1999	Miscellaneous Communication to Applicant - No Action Count
04-06-1998	Information Disclosure Statement (IDS) Filed
04-06-1998	Information Disclosure Statement (IDS) Filed
07-17-1998	Recordation of Patent Grant Mailed
07-13-1998	Sequence Moved to Public Database
06-08-1998	Issue Notification Mailed
05-22-1998	Sequence Forwarded to Pubs on Tape
03-24-1998	Issue Fee Payment Verified
05-12-1998	Drawing(s) Processing Completed
05-08-1998	Drawing(s) Matched to Application
04-15-1998	Drawing(s) Received at Publications
03-24-1998	Mailroom Date of Drawing(s)
01-09-1998	Mail Corrected Notice of Allowance
01-09-1998	Corrected Notice of Allowance
01-08-1998	Communication - Re: Power of Attorney (PTOL-308)
12-24-1997	Correspondence Address Change
12-24-1997	Change in Power of Attorney (May Include Associate POA)
01-08-1998	Mail Notice of Allowance
01-08-1998	Notice of Allowance Data Verification Completed
01-07-1998	Examiner Interview Summary Record (PTOL - 413)
01-06-1998	Terminal Disclaimer Approved in TC
12-24-1997	Terminal Disclaimer Filed
01-05-1998	Error(s) in CRF Corrected by STIC
12-24-1997	Notice of Appeal Filed
01-03-1998	Date Forwarded to Examiner
12-24-1997	Amendment after Final Rejection
01-03-1998	CRF Disk Has Been Received by Preexam / Group / PCT
12-24-1997	Request for Extension of Time - Granted
06-20-1997	Mail Final Rejection (PTOL - 326)
06-19-1997	Final Rejection
04-08-1997	Date Forwarded to Examiner
02-03-1997	Response after Non-Final Action
03-13-1997	CRF Does Not Match Application Specification -- Applicant Must Correct
03-07-1997	CRF Disk Has Been Received by Preexam / Group / PCT
02-26-1997	Case Docketed to Examiner in GAU
01-03-1997	Request for Extension of Time - Granted
08-01-1996	Mail Non-Final Rejection

07-31-1996	Non-Final Rejection
02-12-1996	Case Docketed to Examiner in GAU
01-25-1996	Application Captured on Microfilm
01-18-1996	Application Is Now Complete
08-14-1995	Notice Mailed--Application Incomplete--Filing Date Assigned
08-04-1995	CRF Is Good Technically / Entered into Database
07-26-1995	CRF Disk Has Been Received by Preexam / Group / PCT

If you need help:

- *Call the Patent Electronic Business Center at (866) 217-9197 (toll free) or e-mail EBC@uspto.gov for specific questions about Patent Application Information Retrieval (PAIR).*
- *Send general questions about USPTO programs to the USPTO Contact Center (UCC) .*
- *If you experience technical difficulties or problems with this application, please report them via e-mail to Electronic Business Support or call 1 800-786-9199.*

EXHIBIT D

UNITED STATES DISTRICT COURT
EASTERN DISTRICT OF NEW YORK

COLD SPRING HARBOR LABORATORY,

Plaintiff,

-against-

ROPES & GRAY LLP and MATTHEW P. VINCENT,

Defendants.

Civil Action No.:

COMPLAINT

(SI)

JURY TRIAL DEMANDED

CV-10 0661

SPATT, J.

TOMLINSON, M

Plaintiff, Cold Spring Harbor Laboratory ("CSHL"), by its attorneys, hereby alleges as follows:

PARTIES

1. CSHL is a New York education corporation chartered by the New York State Department of Education, with its principal place of business located in Cold Spring Harbor, New York.

2. Upon information and belief, defendant Ropes & Gray LLP ("R&G") is a Delaware limited liability partnership with its principal place of business in Boston, Massachusetts.

3. Defendant Matthew P. Vincent ("Vincent") was, until April 2009, a registered patent attorney and a partner in R&G's Intellectual Property group. Vincent received a Bachelors of Science degree in 1986 from Worcester Polytechnic Institute, a Ph.D. in Biochemical Sciences in 1991 from Tufts University School of Medicine, and a law degree in

1996 from Suffolk University Law School in Boston. Upon information and belief, Vincent is a resident of the State of Massachusetts.

4. As discussed in greater detail below, on or about July 20, 2009, Vincent filed his resignation from the practice of law with the Office of the Bar Counsel for the State of Massachusetts. Vincent's resignation resulted from an investigation into his misconduct that was the subject of a disciplinary investigation revolving around his having, for more than six years, under cover of a separate company which he formed, billed and collected from R&G's clients more than $700,000 for work that could not be substantiated or verified, purportedly without R&G's knowledge that he owned said company. In late April 2009, R&G terminated Vincent's employment, purportedly because of the misconduct described above.

JURISDICTION AND VENUE

5. This Court has subject matter jurisdiction over this action pursuant to 28 U.S.C. § 1338(a).

6. Venue is appropriate in this judicial district pursuant to 28 U.S.C. § 1391(b)(1) in that R&G is subject to personal jurisdiction in this district and, thus, is a resident of this district, and pursuant to § 1391(b)(2) in that a substantial part of the events or omissions giving rise to CSHL's claims occurred in, and a substantial part of the property that is the subject of the action is situated in, this district.

SUMMARY OF CLAIMS

7. This case involves the negligence of R&G in prosecuting a series of patent applications on behalf of CSHL. These patent applications are directed to inventions made by Dr. Gregory Hannon, a Professor and Howard Hughes Medical Institute Investigator at CSHL, and his colleagues at CSHL (collectively, "Dr. Hannon"), which exploit a cellular mechanism

called RNA interference ("RNAi"). Generally, RNAi refers to the process by which double stranded RNA functions to help regulate when genes are turned off and on in the cell. Through his work at CSHL, Dr. Hannon developed novel methods and technologies to exploit RNAi as a research tool in mammalian cells, which Dr. Hannon achieved by engineering RNA molecules called short hairpin RNAs (shRNAs). Among its many applications, the shRNA technology Dr. Hannon invented gives researchers the ability to specifically turn off expression of virtually any target gene or combination of target genes in a mammalian cell. This valuable technology provides an efficient, effective, widely applicable alternative to more expensive and laborious methods for research and drug development. Today, shRNA has become a fundamental tool in biomedical research for studying what genes do in cells, what goes wrong in diseases such as cancer and for identifying drug targets.

8. In its prosecution of patent applications intended to cover Dr. Hannon's shRNA inventions (the "Hannon Applications"), R&G, and in particular, Vincent, the R&G attorney responsible for prosecution of these patents, committed malpractice through their failure to conduct the prosecution according to a reasonable standard of care, with the result that their conduct has both delayed and prejudiced CSHL's efforts to obtain patent claims covering Dr. Hannon's inventions in several ways, as summarized below.

9. When Vincent drafted the three earliest non-provisional Hannon Applications (U.S. patent applications nos. 09/858,862, filed May 16, 2001, 09/866,557, filed March 24, 2001 and international patent application PCT/US01/08345, filed March 16, 2001), rather than providing an original, complete description of Dr. Hannon's work, Vincent instead relied upon copying extensive portions of text -- essentially verbatim -- from a prior patent application (WO/99/32619) published by a team led by another researcher in the RNAi field, Dr.

Andrew Fire (collectively, "Fire"), to at least, in part, describe Dr. Hannon's inventions. About one half of the "Detailed Description of Certain Preferred Embodiments" found in the three earliest filed Hannon Applications consists of text copied from the Fire application. As described below, Vincent continued to rely upon this text to describe Dr. Hannon's inventions, and in particular, the shRNA technology that is the subject of the pending Hannon Applications. By relying extensively on the copied text, Vincent failed to fully describe and distinguish Dr. Hannon's inventions from the different technology invented by Fire.

10. During the course of prosecution, Vincent and R&G filed numerous follow-up continuation and continuation-in-part ("CIP") applications, allowing several opportunities to properly re-draft the Hannon Applications in such a way that relied on an original description of Dr. Hannon's own work to accurately describe the shRNA technology that Dr. Hannon invented. Instead, R&G and Vincent continued to rely upon the text copied from the Fire application, which falsely implied that Dr. Hannon's shRNA technology was either something that Fire invented or was suggested by the Fire application.

11. R&G and Vincent compounded the prejudice resulting from their continued use of the copied Fire text by failing to properly supplement the foregoing CIP applications with Dr. Hannon's ongoing work in a timely fashion. Furthermore, neither R&G nor Vincent ever brought the fact of this copying of the Fire text, and the potential prejudice resulting from that copying, to the attention of Dr. Hannon and CSHL. Despite being aware of how their negligent conduct had compromised the Hannon Applications, R&G and Vincent continued to prosecute the applications while hiding this fact from Dr. Hannon and CSHL. Vincent and R&G further compounded their negligence by their failure to properly attribute the copied text to Fire, effectively continuing to misrepresent that disclosure as being part of Dr.

Hannon's work, and by years of delay in disclosing the Fire application to the United States Patent & Trademark Office (the "PTO"). Significantly, all this deprived Dr. Hannon and CSHL of any opportunity to mitigate the harm caused by Vincent and R&G's negligent prosecution, including the opportunity to appropriately re-draft the specification in a timely fashion to minimize any potential loss of priority.

12. It was the PTO that first noted the similarity of certain text in the Hannon Applications to that in the Fire application, a fact which ultimately brought R&G's and Vincent's misconduct to CSHL's attention. That fact, however, came to CSHL's attention only after Vincent and R&G had already caused irreparable harm through their negligent prosecution of the Hannon Applications. By the time the PTO eventually cited the Fire application as prior art against Dr. Hannon's invention, the Hannon Applications had been unfairly prejudiced by the erroneous perception that the technology invented by Dr. Hannon is not sufficiently unique from what the Fire application describes to warrant a patent.

STATEMENT OF FACTS

Background and Description of Relevant Technologies

13. CSHL is a private, not-for-profit research and education institution at the forefront of efforts in molecular biology and genetics to generate knowledge that will yield better diagnostics and treatments for cancer, neurological diseases and other major causes of human suffering.

14. Home to eight Nobelists, CSHL was founded in 1890 as one of the first institutions to specialize in genetics research and subsequently has played a central role in the seminal field of molecular biology. At CSHL in 1953, James D. Watson presented his first public lecture on his and Francis Crick's discovery of the double-helical structure of DNA, for

which each later won a Nobel Prize. As Director and then President of the Laboratory from 1968 to 2003, Watson was instrumental in developing CSHL into one of the world's most influential cancer research centers.

15. Today, more than 400 scientists at CSHL pioneer the frontiers of biomedical research. A designated Center of the National Cancer Institute, CSHL has broken new ground in the study of cancer genetics. It has also taken a leading role in efforts to understand what causes neurodevelopmental and neurodegenerative illnesses such as autism, schizophrenia, and Alzheimer's and Parkinson's diseases, and is a global leader in plant genetics and in the emerging discipline of quantitative biology.

16. Each year 8,000 of the world's leading life scientists are drawn to the campus for CSHL's legendary Meetings and Courses program, where new research is discussed and debated. The CSHL Press publishes textbooks and research journals that are among the most highly cited in their fields. CSHL also has created the DNA Learning Center, the nation's first science center dedicated to public genetics education. Its hands-on programs have reached 325,000 middle and high school students, teachers, and families since 1988, and its award-winning website millions more.

17. With regard to the Hannon Applications, of particular importance are the methods and technologies Dr. Hannon invented to use shRNAs in human and other mammalian cells. Since Dr. Hannon's invention, use of shRNA for gene silencing and regulation has become a valuable and widely adopted technology, which is used today in many different fields of medical and pharmaceutical research.

18. In 2002, Dr. Hannon's research on RNA interference was recognized by Science magazine as the Breakthrough of the Year and in 2005 by Esquire as a Breakthrough of

the Decade. Recognized as one of the world's most accomplished scientists, Dr. Hannon has received numerous awards, including appointment as a Pew Scholar in the Biomedical Sciences and as a Rita Allen Foundation Scholar. In 2003, he received the U.S. Army Breast Cancer Research Program's Innovator Award; in 2005 the American Association for Cancer Research's Award for Outstanding Achievement in Cancer Research and in 2007 he received the National Academy of Sciences Award for Molecular Biology and The Memorial Sloan-Kettering Cancer Center's Paul Marks Prize for Cancer Research. He assumed his current position in 2005 as a Howard Hughes Medical Institute Professor and continues to explore the mechanisms and regulation of RNA interference as well as its applications to cancer research.

19. CSHL is the assignee of the entire right, title, and interest in the Hannon Applications, which collectively refer to U.S. patent application numbers 09/858,862 filed May 16, 2001 ("the '862 application"), 09/866,557, filed March 24, 2001 ("the '557 application), 10/055,797 filed January 22, 2002 ("the '797 application"), 10/350,798 filed January 24, 2003, 10/997,086, filed November 23, 2004, 11/791,554 filed May 23, 2007, 11/894,676 filed August 20, 2007, 12/152,655 filed May 15, 2008, 12/152,837 filed May 16, 2008 and international patent applications PCT/US01/08435 filed March 16, 2001 ("the '435 PCT application"), PCT/US03/01963 filed January 22, 2003, and PCT/US05/42488 filed November 23, 2005, including all foreign patent applications filed therefrom. Certain of the Hannon Applications claim a benefit of priority to U.S. provisional applications 60/189,739 filed March 16, 2000 ("the '739 application"), and 60/243,097 filed October 24, 2000 ("the '097 application).

20. International Patent Application PCT/US98/27233, which was published on July 1, 1999 with International Publication Number WO/99/32619 (the "Fire application"), describes certain work conducted by Fire relating to his discovery that long double stranded

RNA molecules could specifically silence gene expression in invertebrate cells. Fire referred generally to this cellular process as RNA interference, or RNAi.

21. Fire received U.S. Patent No. 6,506,559 for his RNAi technology, which issued on January 14, 2003 (the "Fire Patent"). The essentially identical written disclosures of the Fire application and the Fire Patent are referred to collectively hereinafter as the "Fire Specification."

Facts Relating to Malpractice

22. From in or around 2001 until late 2008, R&G acted as principal outside patent prosecution counsel for CSHL.

23. Vincent was the R&G attorney primarily involved in the drafting and prosecution of the all of the Hannon Applications, as well as the '739 application and the '097 application, to which certain of the Hannon Applications claim priority. To date, CSHL has paid R&G approximately $420,000 in legal fees and disbursements that R&G has billed for its prosecution of the patent applications related to Dr. Hannon's shRNA technologies, and approximately $1,400,000 in fees and disbursements that R&G has billed for its prosecution of other applications.

24. The '097 application includes about 11 pages of text that Vincent copied essentially verbatim, without citation or attribution, from the published Fire application. Attached as Exhibit A is a "marked-up" version of the '097 application, in which text that is the same as text in the Fire Specification is highlighted. Vincent specifically carried over at least some portion of the copied Fire text found in the '097 application into all of the subsequently filed Hannon Applications.

25. Despite the fact that Vincent was fully aware of the Fire application as early as October 2000, when he first copied text from that application into the '097 application, R&G did not cite the Fire application in any papers filed on behalf of CSHL until the Supplemental Information Disclosure Statement of November 26, 2004. And, R&G did not cite the Fire Patent in any papers filed on behalf of CSHL until the Supplemental Information Disclosure Statement of January 7, 2005.

26. The '435 PCT application, filed March 16, 2001, the '862 application filed May 16, 2001, and the '557 application filed May 24, 2001 are directed generally to the initial methods and technologies Dr. Hannon developed relating to use of RNA interference in mammalian and other cells, including use of hairpin RNAs to regulate target genes. In particular, the filed '557 application included claims (20-25) directed to use of hairpin RNAs to inhibit gene expression and expression of such hairpin RNAs in cells of a transgenic non-human mammal.

27. About one half of the "Detailed Description of Certain Preferred Embodiments" (hereinafter, the "Detailed Description") found in the three earliest filed non-provisional Hannon Applications consists of text copied from the Fire application. Vincent's failure to provide an adequate description of Dr. Hannon's technology in these applications seriously compromised the ability of these applications, in particular the '557 application, to serve as priority support for Dr. Hannon's patent claims. This fact has deprived CSHL of the opportunity to obtain allowance of claims covering Dr. Hannon's inventions entitled to the respective filing dates of these applications, based on the support from these applications.

28. Vincent repeatedly effectuated this copying of Fire, notwithstanding that he knew from the outset that the Hannon Applications needed to be distinguished from Fire.

29. In this regard, the now pending claims of the Hannon '086 application are directed to one particularly valuable aspect of the technology Dr. Hannon developed, methods that allow one to stably suppress gene expression in mammalian cells using RNA interference. Among other things, this valuable technology provides an efficient, effective and widely applicable alternative to more expensive and laborious methods for biomedical research and drug development. Dr. Hannon's shRNA methods represented a considerable advance over the prior art, including the Fire patent, which failed to provide any solution for how to use RNA interference in mammals, without killing the treated cells through the so-called "interferon" or "protein kinase (PK) response".

30. Vincent first added claims directed to use of hairpin RNAs to inhibit gene expression and expression of such hairpin RNAs in cells of a transgenic non-human mammal in the '557 application, which was filed May 24, 2001 as a continuation-in-part ("CIP") of the '435 PCT application. Both from the standpoint of meeting his duty of care and scientifically, in adding these new claims, Vincent had an affirmative duty to amend the specification with original text accurately describing these additional claimed inventions and distinguish these from the Fire application. However, instead of properly amending the Detailed Description, Vincent again relied on the same text he had previously copied from the Fire application as support for these new claims, knowing full well that text copied directly from Fire could not serve to distinguish the newly claimed subject matter from Fire.

31. During 2001, Vincent regularly communicated with Dr. Hannon regarding the Hannon Applications Vincent was then prosecuting. In filing the '557 application on May 24, 2001, and also in filing the '797 application in January 22, 2002, Vincent either knew or should have known the relevance and potential prejudice of continuing to rely on extensive

passages of text copied from the Fire application. He should have told Dr. Hannon then what he had done. However, before filing the '557 application, and even the '797 application, Vincent failed to inform Dr. Hannon and CSHL (either directly or through Dr. Hannon) of this copying of Fire and Vincent's continued use of this copied text in the Detailed Description as support for the filed claims. His failure to do so deprived Dr. Hannon and CSHL of the timely opportunity to amend the specification to properly describe and distinguish Dr. Hannon's technology from the different methods Fire described.

32. As corroborated by laboratory and other records, by the time Vincent had filed the '557 application, Dr. Hannon had already conceived of the short hairpin methods that are the subject of the now pending Hannon claims. Had Vincent in May 2001 informed CSHL and Dr. Hannon of his conduct, such information would have identified an urgent need to amend the '557 application to distinguish Dr. Hannon's hairpin technology, including in particular the use of short hairpins, from Fire's altogether different disclosure. Such amendment would necessarily have included adding original disclosure describing Dr. Hannon's short hairpin invention.

33. Vincent's failure to inform CSHL and Dr. Hannon of his conduct resulted in an entirely unnecessary and prejudicial delay in adding specific disclosure about the short hairpin invention to the Hannon Applications. To the extent Vincent eventually did so, this happened eight months later with the filing of the '797 CIP application on January 22, 2002. Even then, instead of revising the Detailed Description to provide an accurate description of the short hairpin technology, Vincent continued to improperly rely on the text he had copied from the Fire application, fully knowing that this text was directed to an entirely different invention. Had Vincent been forthright in May of 2001 about his copying of the Fire text, Vincent would

have no doubt been apprised then (if he was not already aware) of Dr. Hannon's work relating to the short hairpin invention. Instead, by waiting until January 2002 to get reference to short hairpins into Dr. Hannon's applications, Vincent caused a potential crucial loss of priority from May 2001.

34. In short, even after the filing of the '557 application, Vincent and R&G failed to comply with a reasonable standard of care in the subsequent prosecution of these applications and the filing of subsequent applications in the PTO based on these parent applications. Vincent and R&G never brought the fact of Vincent's copying of the Fire text, and the potential prejudice resulting from that copying, to the attention of CSHL. Despite being aware of how his conduct had compromised the Hannon Applications, Vincent continued to prosecute them while hiding this fact from CSHL. Significantly, this deprived CSHL of any opportunity to address the issues the copying raised early in prosecution, when the applications could have been re-drafted in a timely fashion to minimize any potential loss of priority and minimize the harm to CSHL.

35. Further compounding the harm Vincent had caused, in prosecuting these early applications, Vincent's and R&G's improperly relied on and misrepresented the copied Fire text as describing the technology Dr. Hannon invented. In effect, their actions erroneously implied that Dr. Hannon's technology was previously invented or described by Fire, which it was not. Through their failure to properly attribute the copied text to Fire, and years of delay in even bringing the Fire application and Fire Patent to the attention of the PTO, Vincent and R&G further compounded their malpractice by effectively continuing to misrepresent that disclosure as being part of Dr. Hannon's work.

36. For example, these misrepresentations include statements made during prosecution of the '557 application. In the office action dated April 21, 2005 rejecting all pending claims, the PTO Examiner argued that the specification failed to teach introducing an expression vector encoding a hairpin RNA into mammalian cells. In response (Reply and Amendment filed August 11, 2005), R&G argued that the application described the use of expression systems that are intended to produce hairpin RNAs upon being transcribed in cells. In support, R&G repeatedly pointed to various sections of text copied from the Fire application. To support its position, R&G filed a Rule 132 Expert Declaration (Declaration under 35 U.S.C. §1.132 of Frank McKeon dated July 29, 2005), which cited repeatedly to sections of the copied Fire text as evidence that the technology invented by Dr. Hannon and described in the '557 application was directed to use of expression systems intended to produce hairpin RNAs upon being transcribed in cells.

37. As stated above, instead of properly re-drafting the Hannon Applications via the numerous CIP applications, in a way that relied upon an original description of Dr. Hannon's work, Vincent and R&G continued to rely on text copied from Fire despite the fact that this risked the false implication that Dr. Hannon's shRNA technology was either something that Fire invented or was suggested by the Fire application.

38. Notably, on August 11, 2005, the PTO issued a Notice of Allowance for the pending '557 application claims. In explaining his reasons for allowance, the Examiner stated on page 3:

> The declarations under 37 CFR 1.132 filed August 11, 2005 are sufficient to overcome the rejections of claims...based upon new matter under 35 USC 112 first paragraph and lack of enablement under 35 USC 112 first paragraph. Specifically, the declaration of Frank Mckeon establishes that the double stranded RNA construct

of the patent application encompass hairpin RNA comprising the features claimed therein. The declaration provides evidence by specific examples in the instant application.

39. On April 6, 2006, however, the PTO withdrew the '557 application from issue to reconsider its decision. On September 6, 2006, the Examiner rejected all pending claims as anticipated by the Fire Patent.

40. The prejudice to the Hannon Applications caused by Vincent's original attempt to describe the Hannon inventions by copying text from Fire application was illustrated by R&G's subsequent failed efforts to overcome the Fire Patent as prior art against the '557 application. In the Amendment filed March 9, 2007, Vincent argued that in contrast to the disclosure of the '557 application, the Fire Patent failed to provide any particular guidance that would have led one to envisage the claimed methods directed to using an expression vector encoding a hairpin RNA to attenuate gene expression specifically in mammalian cells.

41. In fact, before filing the March 9, 2007 Amendment, Vincent conducted an in-person interview in February with Examiner McGarry and Supervisory Examiner Schultz at the PTO, specifically to discuss the Examiner McGarry's rejection of the pending claims as anticipated by the Fire Patent. Notably, Vincent brought with him to the interview both Dr. Hannon and John Maroney, the Vice President, Legal Counsel, and Director of CSHL's Office of Technology Transfer. During this interview, Vincent provided the Examiners with a preview of the argument he planned to present in the March 9, 2007 Amendment. Despite the fact that his planned arguments relied on text copied from the Fire specification, Vincent never disclosed this fact to Examiners McGarry and Schultz, Dr. Hannon or Mr. Maroney.

42. In the Office Action dated September 4, 2007, the PTO rejected Vincent's argument that only the '557 specification provided such guidance. Referring specifically to the text that Vincent had copied from Fire, the PTO noted that "in fact the disclosure of cell/organisms of the instant specification at pages 21-22 is essentially verbatim of the disclosure of Fire et al at column 8," and that "it is unclear how applicant claimed invention differs from what has been disclosed by the prior art."

43. The '797 application was filed as a continuation-in-part of the '435 PCT application and among other things, incorporated additional disclosures from Dr. Hannon relating to the use of the shRNA technology for regulating gene expression in mammalian cells. This added material was directed to an entirely different invention from the technology described in the Fire application. Despite that fact, Vincent retained the copied Fire text in the Detailed Description, fully knowing that the copied Fire text described a different invention.

44. In copying Fire again in filing the '797 application, Vincent furthered his malpractice by again failing to make any reasonable effort to amend the specification, in either the summary or detailed description of the invention, to accurately support claims directed to use of short hairpins in mammalian cells. Instead, Vincent continued to rely on the extensively copied Fire text as support. The prejudicial consequence of Vincent's actions has been repeatedly demonstrated in attempts to draft claims specifically defining "short hairpins" in a manner that unambiguously distinguishes "short hairpins" from the "long hairpins" that the PTO (albeit improperly) now alleges are described by Fire.

45. In this regard, during prosecution of the '557 application, the examiner rejected Vincent's attempt to add a numerical limitation to the hairpin claims, noting that the attempt to add a specific numeric upper limit to such claims "is not consistent with the

specification and constitutes new matter where it is not disclosed nor made apparent by the disclosure of the specification that such a specific range was intended."

46. With respect to the short hairpin claims now pending in the '086 and '676 applications, Vincent's failure to properly supplement the '797 application with original disclosure specifically describing the short hairpin invention compounded the problem Vincent created by improperly relying on the copied Fire text. The continued harm these actions has caused is evident from the examiner's pending rejection of short hairpin claims in the '086 and '676 applications (filed from and claiming priority to the '797 application) as being anticipated by Fire.

47. Vincent's failure to properly distinguish Dr. Hannon's invention was further compounded by his failure to include additional disclosure regarding use of shRNA in mammalian cells that Dr. Hannon had available at the time that he filed the '797 application. This material, among other things, included data and other information Dr. Hannon provided in the article published as Genes & Development 16:948-958, a draft of which Dr. Hannon first sent to the journal on or about the '797 filing date.

48. In short, as a direct result of Vincent's and R&G's malpractice, the Hannon Applications have been unfairly prejudiced and compromised by the erroneous perception that the technology invented by Dr. Hannon is not sufficiently unique from what the Fire application describes to warrant a patent. Such a perception, created by Vincent's and R&G's misconduct, directly contributed to R&G's failure to obtain any allowed claims covering Dr. Hannon's invention.

49. Before, and increasingly throughout 2007, Mr. Maroney experienced frustration with a lack of communication from Vincent about the Hannon Applications, and in

particular about Vincent's understanding of why the PTO had withdrawn the '557 application from issue.

50. Dr. Vladimir Drozdoff is a Senior Licensing Associate and Patent Attorney for CSHL's Office of Technology Transfer. Dr. Drozdoff first began working at CSHL on February 11, 2008, with one of his first priorities being a review of Vincent's prosecution of the Hannon Applications.

51. Dr. Drozdoff first became aware of apparent copying of text from the Fire Specifications into the Hannon Applications upon reviewing the Office Action of September 4, 2007. In that Office Action, the Examiner noted the similarity of certain text to that in the Fire specification, in particular on both page 8 and page 10, that "*the disclosure of cell[s]/organisms of the instant specification at pages 21-22 is essentially verbatim of the disclosure of Fire et al. at column 8.*"

52. Dr. Drozdoff then compared the entire '557 specification with the entire Fire Patent, observing that the similarity extended beyond column 8 of the Fire Patent. Consequently, he conducted a further review of the Hannon Applications, which included a detailed comparison of the '739 application and the '097 application, to which the Hannon Applications claim priority, to the text of the Fire Patent and the published Fire application.

53. As a result of Dr. Drozdoff's investigation, the following facts were evident: (1) The "Summary of the Invention" of the '097 application contains approximately 11 pages of text that is almost identical to text found in the Fire application and in the Fire Patent; (2) this text is found in the Fire application in certain portions of 3 pages within the "Summary of Invention" and in certain portions of 13 pages within the "Detailed Description of the Invention"; (3) a substantial portion of this text was carried forward into the specification of the various

Hannon Applications; (4) none of the sections of the Hannon Applications where this text appears cite or reference the Fire application; and (5) the Fire application was first disclosed in the '557 prosecution in a November, 2004 filed IDS.

54. Mr. Maroney first became aware of apparent copying of text from the Fire Specifications into the Hannon Applications upon being advised of the same by Dr. Drozdoff on March 3, 2008. Before Dr. Drozdoff provided Mr. Maroney with this information on March 3, 2008, it had been his belief that the text of the '097 priority application and of the various Hannon Applications consisted of text drafted by Vincent along with text generated by the inventors, such as portions of draft manuscripts. This was the first time Mr. Maroney, or, to the best of his knowledge, anyone at CSHL, had been informed or had become aware that either the '097 priority application or any of the Hannon Applications contained numerous pages of text almost identical to text appearing in the Fire application.

55. Dr. Hannon first became aware of apparent copying of text from the Fire Specifications into the Hannon Applications upon being advised of the same by Dr. Drozdoff and Mr. Maroney on March 18, 2008.

56. To the extent that any papers were filed with the PTO, or any statements were made to the PTO, during the prosecution of the Hannon Applications, including any statements made to the PTO about the Fire Patent or the Fire application or any statements made about the Hannon Applications that involved sections that are the same as sections of the Fire Specifications, all such statements were made without any knowledge on the part of Dr. Drozdoff, Mr. Maroney, or Dr. Hannon, or, to the best of their knowledge, on the part of CSHL, that the specifications of the Hannon Applications contain text that is the same as, or very similar to, text from the Fire Specifications.

57. Upon being made aware of the above facts, in accordance with CSHL's duty of candor and good faith in dealing with the PTO, and with CSHL's desire to maintain the highest levels of scientific integrity, Dr. Drozdoff and Mr. Maroney, acting on behalf of CSHL, took steps to further investigate and diligently rectify any impropriety that may have occurred in the prosecution of the Hannon Applications, as a result of the facts described above. These steps included meeting with Vincent and R&G partner James Haley to inform them of CSHL's findings, requesting R&G's full cooperation to provide the PTO with complete disclosure of these facts and in carrying out an entire, orderly transfer of responsibility for prosecution of the Hannon Applications from Vincent.

58. Accordingly, to understand what Vincent knew, Dr. Drozdoff and Mr. Maroney arranged to meet in New York City with Vincent and James Haley, a partner and patent attorney at R&G. During this meeting, which was held on April 1, 2008, Dr. Drozdoff and Mr. Maroney informed them of Dr. Drozdoff's findings regarding the similarity of certain text in the Hannon Applications and the Fire application and Fire Patent, and provided Vincent and Mr. Haley with a comparison of the '097 application to the Fire Specifications. Vincent acknowledged that in filing the '097 application, he was aware that portions thereof had been copied from the published Fire Specifications. At no time during this meeting did Vincent indicate that he had ever informed the PTO, CSHL or any of the co-inventors of this fact.

59. R&G refused to assure CSHL that it would unconditionally assist CSHL in providing the PTO with all relevant facts, instead imposing on CSHL the precondition that required CSHL to first sign a waiver essentially releasing R&G from any liability for misconduct. So that CSHL could carry out its first priority to provide the PTO with full disclosure, CSHL engaged Wilmer Cutler Pickering Hale and Dorr, LLP as new counsel.

60. As a result of their malpractice, R&G and Vincent failed to obtain allowance of any of the Hannon Applications and have caused CSHL to incur substantial, unnecessary and avoidable costs for their prosecution. Not only has CSHL been damaged in the form of hundreds of thousands of dollars in legal fees paid to R&G that would not have been necessary had Vincent and R&G met their duties of care, but CSHL, having been denied patents that it otherwise would have received but for R&G's and Vincent's negligence, has lost millions of dollars in potential licensing fees for the Hannon Applications.

Vincent's Termination by R&G and Subsequent Resignation from the Practice of Law

61. In late April 2009, R&G abruptly terminated Vincent's employment.

62. The purported basis for this termination is conduct by Vincent that was the subject of a pending State of Massachusetts disciplinary proceeding at the time of Vincent's termination, as described below.

63. On or about July 20, 2009, Vincent tendered his voluntary resignation from the practice of law, as a result of said investigation. Vincent's own affidavit in support thereof acknowledged the truth of the material facts upon which the disciplinary charges were based, including the following:

a. At some time prior to April 2002, Vincent formed a business entity known as "The IP Resource Company" to perform patent database searches;

b. Vincent did not inform his partners at R&G or his clients that he was the owner and operator of The IP Resource Company;

c. Beginning in approximately April 2002 and continuing through approximately September, 2008, Vincent prepared and submitted to R&G for payment sixty separate invoices from The IP Resource Company, each invoice relating to multiple patent matters;

d. The invoices that Vincent prepared stated, in summary form, that The IP Resource Company had performed research tasks on a total of approximately 3449 separate client matters and was entitled to payment of a total of $733,771.30 for those services. The invoices did not itemize costs, services rendered, dates on which services were rendered, or time spent.

e. Vincent approved each of the sixty invoices for payment and forwarded them to R&G's accounting department.

f. Relying on Vincent's approval, R&G paid the invoices and billed the appropriate clients for the service.

g. Vincent endorsed the checks for deposit and caused them to be deposited in an account for his personal use.

h. Vincent either never maintained or did not retain the underlying billing records for the invoices submitted by the IP Resource Company, and he cannot satisfactorily account for costs incurred and services rendered.

64. CSHL was among the clients of R&G who were victimized by Vincent's conduct described above. While serving as CSHL's primary patent prosecution counsel, R&G

billed and collected approximately $10,000 from CSHL in the name of work allegedly performed by IP Resource Company at the request, and/or with the approval, of R&G.

FIRST CLAIM
(Legal Malpractice)

65. CSHL repeats and realleges the allegations made in paragraphs 1 through 64 above.

66. Vincent and R&G acted as attorneys for CSHL from in or around 2001 through September 2008. During this time frame, Vincent and R&G owed CSHL a duty of care.

67. During this time frame, Vincent and R&G were the only attorneys acting on behalf of CSHL in the patent prosecution of the Hannon Applications.

68. In taking the actions, and omitting to act, as described in Paragraphs 22-60 above, Mr. Vincent's and R&G's actions were negligent and fell below the applicable professional standards for representing an entity in CSHL's position in pursuing patents such as the Hannon Applications.

69. But for Vincent's and R&G's failures to satisfy their duties of care to CSHL, it would have been able to obtain allowance of claims covering Dr. Hannon's inventions.

70. As a direct and proximate result of Vincent's and R&G's negligence, CSHL has been severely damaged. CSHL has spent hundreds of thousands of dollars on legal fees that would not have been necessary had Vincent and R&G met their duties of care. These fees include substantial additional costs that CSHL has incurred resulting from its transfer of prosecution to new counsel and in legal work that new counsel has conducted in an attempt to address the harm caused by R&G's misconduct to the Hannon Applications, as well as fees paid to R&G that would not have been necessary had Vincent and R&G satisfied their duties of care.

CSHL will establish the exact amount of this category of damages at trial, but expects it to be no less than $1,000,000.

71. In addition, CSHL has lost opportunities for licensing Dr. Hannon's technology, including at minimum, lost opportunities resulting from the loss of patent term caused directly by Vincent's and R&G's malpractice. The resulting losses include lost opportunities for commercial user licenses, allowing commercial use of shRNA technology, as well as lost royalty income on research reagent sales. The exact amount of such damages will be established at trial. However, CSHL expects that the annual amount of such licensing opportunities it has lost as a direct and proximate result of Vincent's and R&G's malpractice to be at least $22,500,000 for lost commercial user license income and $9,000,000 for lost royalty income, should CSHL ultimately obtain the patent such that only five years of patent term are lost, to $57,500,000 for lost commercial user license income and $19,000,000 for lost royalty income, should CSHL be denied the patent such that fifteen years of patent term are lost.

72. In addition, as a direct and proximate result of their negligence, Vincent's and R&G's failure to obtain issuance of claims to CSHL's shRNA technology has precluded CSHL from pursuing opportunities to further commercialize this technology through a start-up company. The exact amount of such damages will be established at trial. However, CSHL expects that the start-up income it has lost as a direct and proximate result of Vincent's and R&G's malpractice to be at least $5,000,000. In sum total, depending on the number of years of term lost, CSHL expects its total damages as a direct and proximate result of Vincent's and R&G's malpractice to be at least $37,500,000 to $82,500,000.

SECOND CLAIM
(Breach of Fiduciary Duty)

73. CSHL repeats and realleges the allegations made in paragraphs 1 through 72 above.

74. As attorneys for CSHL, R&G and Vincent owed CSHL a fiduciary duty.

75. In taking the actions, and omitting to act, as described in Paragraphs 22-60 above, Mr. Vincent's and R&G's actions were in breach of their fiduciary duties to CSHL.

76. As a direct and proximate result of R&G's and Vincent's breaches of their fiduciary duties to CSHL, as described above, CSHL has been damaged in an amount to be determined at trial.

77. Specifically, CSHL has been damaged in the form of: (i) additional attorneys' fees that it has been forced to expend as a result of R&G's and Vincent's wrongful acts and omissions, which is estimated to be no less than $500,000; (ii) lost licensing opportunities for the Hannon technology, which is estimated to be worth no less than $36,500,000 to $81,500,000; and (iii) disgorgement of all attorneys' fees paid by CSHL to R&G since 2001, which is estimated to be no less than $1,400,000.

THIRD CLAIM
(Fraud and Fraudulent Concealment)

78. CSHL repeats and realleges the allegations made in paragraphs 1 through 77 above.

79. For more than eight years, while acting as primary patent prosecution counsel for CSHL, by intentionally and repeatedly concealing Vincent's copying of Fire, Vincent and R&G committed a fraud upon CSHL.

80. Vincent and R&G failed to inform CSHL of the copying of Fire, and failed to inform the PTO of the copying of Fire, despite their professional obligation to make

such disclosure. This withholding of material information was intentionally done by Vincent and R&G for the purpose of, and had the effect of, inducing CSHL to continue using Vincent and R&G as its primary patent counsel, and to reasonably rely on their advice in the mistaken belief that they had undertaken to provide professional advice in CSHL's best interests and had done nothing to stain CSHL's reputation or credibility with the PTO, or prejudice its patent applications.

81. On this point, shortly before filing the March 9, 2007 Amendment in the '557 application, Vincent went with Dr. Hannon and Mr. Maroney to meet with the examiner, essentially over the examiner's pending rejection of claims as being anticipated by Fire. Incredibly, Vincent still never brought to their attention the copying and resulting problems that he created in distinguishing Dr. Hannon's work from Fire.

82. Had Vincent been forthright about his copying of Fire anytime during the course of the prosecution of the Hannon Applications, CSHL would never have kept Vincent and R&G as its patent counsel. By intentionally concealing what actually had transpired, Vincent and R&G intentionally induced CSHL to continue using Vincent and R&G for a majority of its patent work.

83. Vincent and R&G, as attorneys for CSHL, had a duty to disclose their copying of Fire and the prejudice to the Hannon Applications that their continued reliance on the copied text had caused, as described above, which was material information.

84. As a result of this fraudulent concealment and CSHL's reasonable reliance on the deceptive statements made by Vincent and R&G (via the filed Hannon Applications) that failed to disclose the copying of Fire, CSHL has been damaged in an amount to be determined at trial, including punitive damages.

85. Vincent committed a separate fraud upon CSHL through his tendering, through R&G, of invoices from The IP Resource Company totaling at least $9587.45, all of which were paid by CSHL, notwithstanding that, upon information and belief, Vincent could not substantiate any work performed by that company for the benefit of CSHL. The statements of work performed set forth on these invoices constituted material misrepresentations.

86. Vincent tendered these invoices to CSHL, through R&G, knowing that they were false, with the intent that CSHL would rely on the statements contained in those invoices, and the fact that they were being tendered to CSHL by R&G, in making payment thereon.

87. CSHL reasonably relied on the statements of work performed contained in the invoices of The IP Resource Company invoices forwarded for payment by R&G.

88. In making payment on these fraudulent invoices, CSHL suffered damages in an amount to be determined at trial, plus punitive damages.

JURY DEMAND

Pursuant to Rule 38 of the Federal Rules of Civil Procedure, CSHL demands a trial by jury of all of its claims.

PRAYER FOR RELIEF

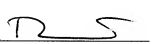
Wherefore, CSHL prays for judgment against R&G and Vincent as follows:

a. Granting R&G and Vincent judgment on its Claims in a sum to be determined at trial, which is expected to be no less than $37,500,000 to $82,500,000 plus punitive damages;

b. Awarding CSHL attorneys fees and costs; and

c. Awarding CSHL such other and further relief as this Court deems just and proper.

Dated: February 16, 2010
Garden City, New York



Chad E. Ziegler [CZ5717] [Pro Hac Vice-pending]
Peter I. Bernstein [PB3549]
Scully, Scott, Murphy & Presser, P.C.
400 Garden City Plaza, Suite 300
Garden City, New York 11530
516-742-4343

Exhibit A

US 20020162126A1

(19) **United States**

(12) **Patent Application Publication** (10) **Pub. No.: US 2002/0162126 A1**

Beach et al. (43) **Pub. Date: Oct. 31, 2002**

(54) **METHODS AND COMPOSITIONS FOR RNA INTERFERENCE**

(76) Inventors: **David Beach**, Boston, MA (US); **Emily Bernstein**, Huntington, NY (US); **Amy Caudy**, Melville, NY (US); **Scott Hammond**, Huntington, NY (US); **Gregory Hannon**, Huntington, NY (US)

Correspondence Address:
ROPES & GRAY
ONE INTERNATIONAL PLACE
BOSTON, MA 02110-2624 (US)

(21) Appl. No.: **09/866,557**

(22) Filed: **May 24, 2001**

Related U.S. Application Data

(63) Continuation-in-part of application No. PCT/US01/08435, filed on Mar. 16, 2001.

(60) Provisional application No. 60/189,739, filed on Mar. 16, 2000. Provisional application No. 60/243,097, filed on Oct. 24, 2000.

Publication Classification

(51) **Int. Cl.**[7] **A01K 67/00**; A61K 48/00; C12N 15/87

(52) **U.S. Cl.** **800/8**; 514/44; 435/455

(57) **ABSTRACT**

The present invention provides methods for attenuating gene expression in a cell using gene-targeted double stranded RNA (dsRNA). The dsRNA contains a nucleotide sequence that hybridizes under physiologic conditions of the cell to the nucleotide sequence of at least a portion of the gene to be inhibited (the "target" gene).

Patent Application Publication Oct. 31, 2002 Sheet 1 of 34 US 2002/0162126 A1

Figure 1

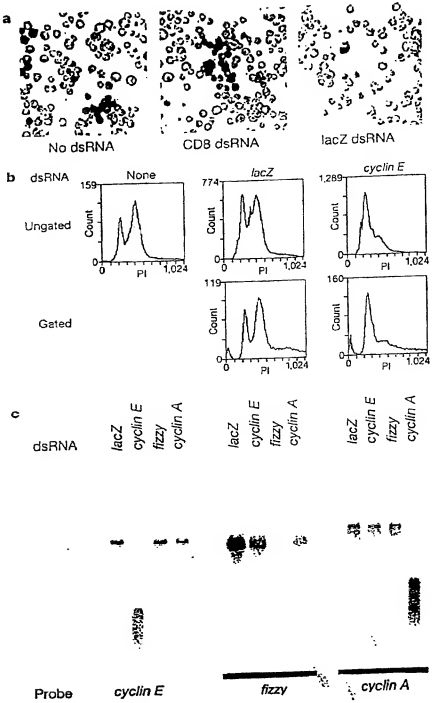


Figure 2

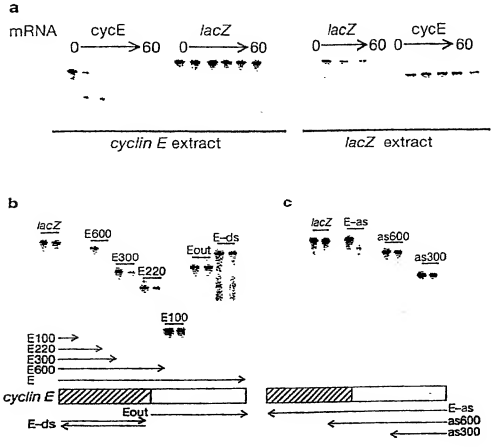

a
mRNA
cycE
lacZ
0
60
lacZ
cycE
cyclin E extract
lacZ extract
b
lacZ
E600
E300
E220
E-ds
Eout
E100
E100
E220
E300
E600
E
cyclin E
Eout
E-ds
c
lacZ
E-as
as600
as300
E-as
as600
as300

Figure 3

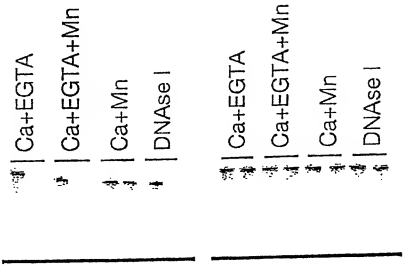

Ca+EGTA
Ca+EGTA+Mn
Ca+Mn
DNAse I
Ca+EGTA
Ca+EGTA+Mn
Ca+Mn
DNAse I
Substrate
cyclin E
lacZ

Figure 4

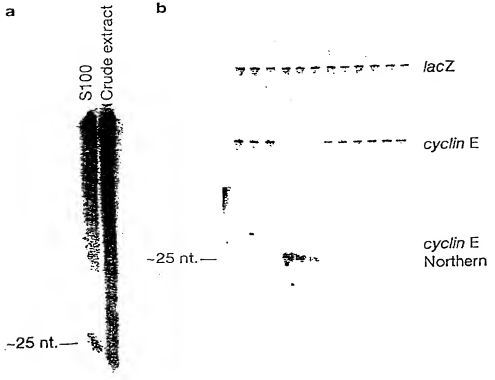

a
S100
Crude extract
~25 nt.—
b
lacZ
cyclin E
cyclin E
Northern
~25 nt.—

Figure 5

Figure 6a-c

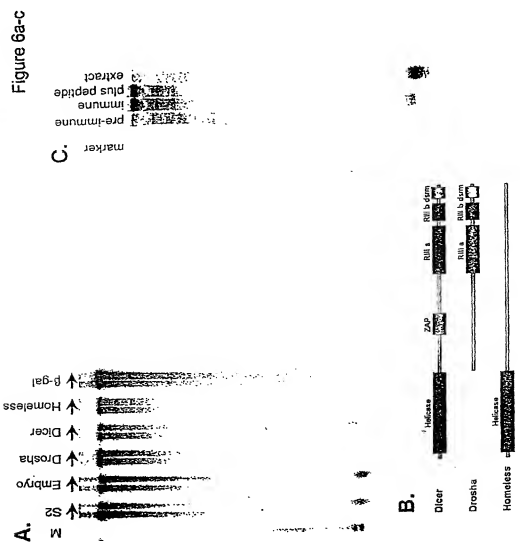
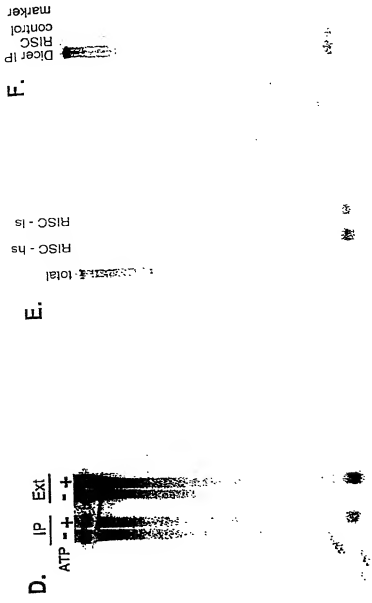

A.
M
S2
Embryo
Drosha
Dicer
Homeless
β-gal
B.
Dicer
Drosha
Homeless
Helicase
ZAP
RIII a
RIII b dsrm
C.
marker
pre-immune
immune
plus peptide
extract

Figure 6d-f


F.
Dicer IP
RISC
control
marker
E.
total
RISC - hs
RISC - ls
D.
ATP
IP
Ext

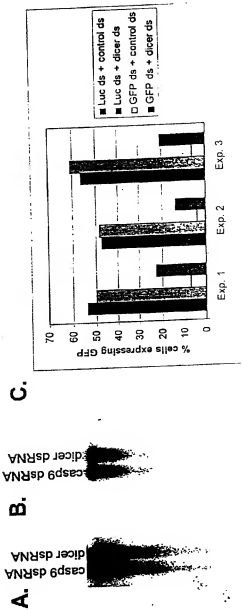

C.
% cells expressing GFP
0
10
20
30
40
50
60
70
Exp. 1
Exp. 2
Exp. 3
Luc ds + control ds
Luc ds + dicer ds
GFP ds + control ds
GFP ds + dicer ds
B.
dicer dsRNA
casp9 dsRNA
A.
dicer dsRNA
casp9 dsRNA

Figure 7

Figure 8A, B

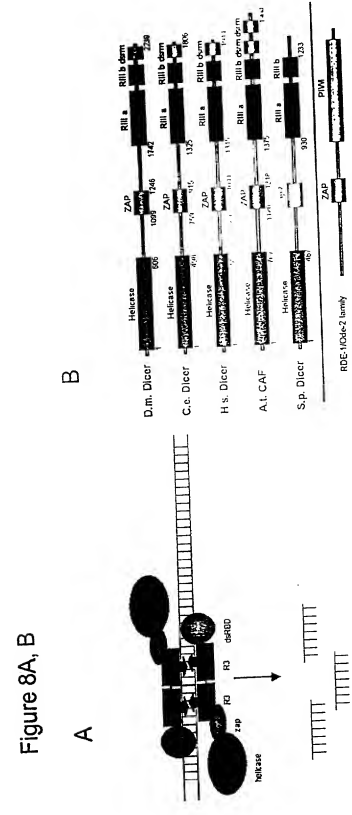
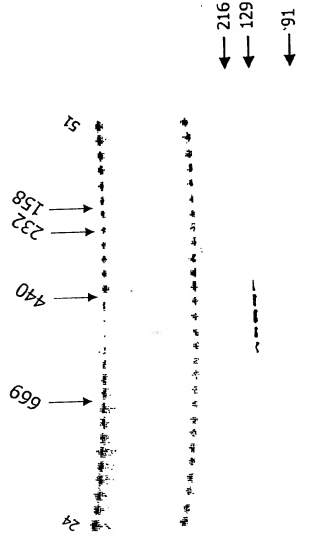

A
B
zap
helicase
R3
dsRBD
D.m. Dicer
C.e. Dicer
H.s. Dicer
A.t. CAF
S.p. Dicer
RDE-1/QDE-2 family
Helicase
ZAP
RIII a
RIII b
dsrm
PIWI

Figure 9


Cell free extract
200K spin
Ribosome pellet
High salt extraction
Ribosome associated proteins
Low salt precipitation
Pellet
Supernatant
RNAi activity
Resolubilize High Salt
Chromatography
Superose 6
Mono S
Mono Q
Hydroxyapatite

Figure 10


Size Std, kD
669
440
232
158
Fraction
54
51
Luciferase transcript
Cyclin E (control) transcript
dArgo-2 immunoblot
216
129
91

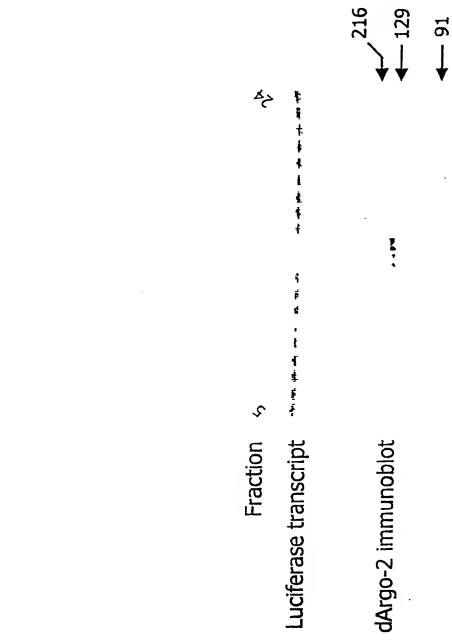

Fraction
Luciferase transcript
dArgo-2 immunoblot
216
129
91

Figure 11

Figure 12

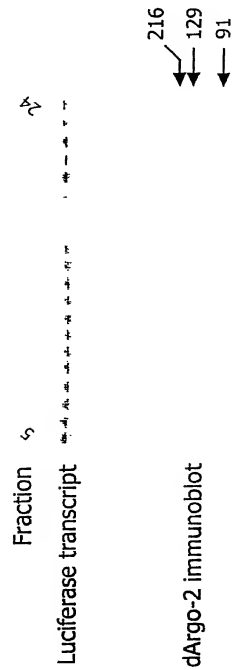
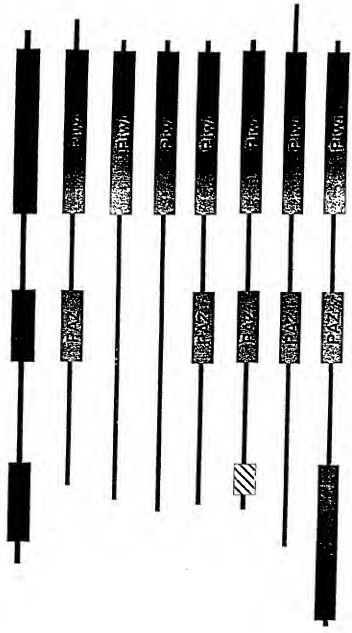

Fraction
5
Luciferase transcript
dArgo-2 immunoblot
216
129
91

Figure 13

Figure 14


At-argonaute-1
Hs- eIF2C
Nc-qde-2
Ce-rde-1
Dm-piwi
Dm-sting
Dm-argonaute-1
Dm-argonaute-2
PAZ
PIWI

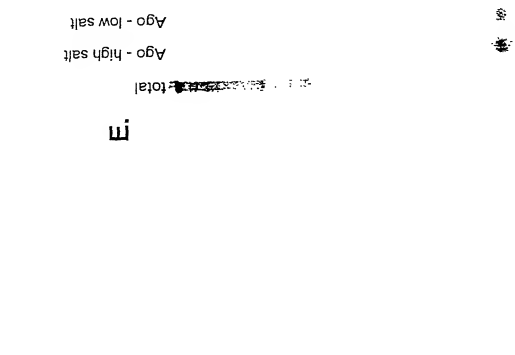
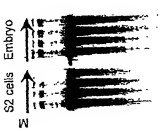


Figure 15

Figure 16



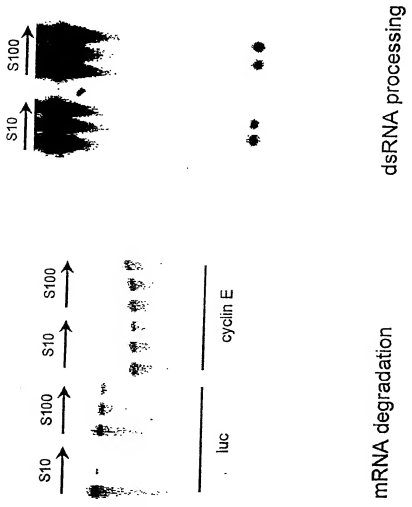

S100
S10
dsRNA processing
S10
S100
S10
S100
luc
cyclin E
mRNA degradation

Figure 17

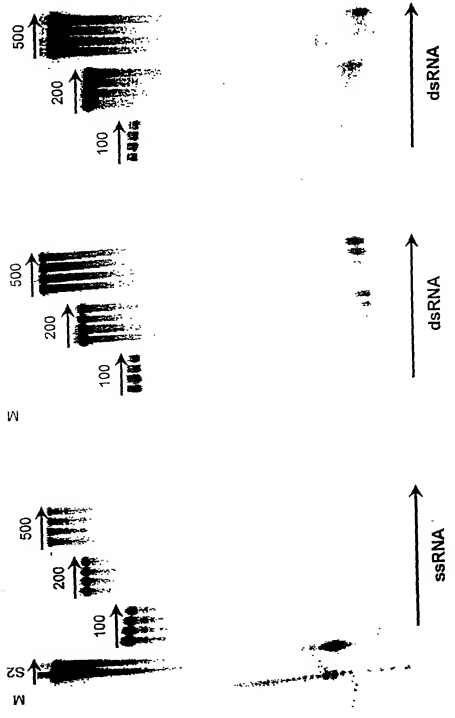


Figure 18

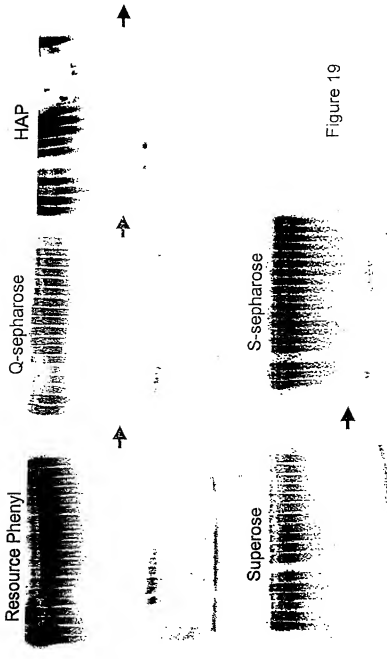


Figure 19

Purification of the 22-mer generating enzyme

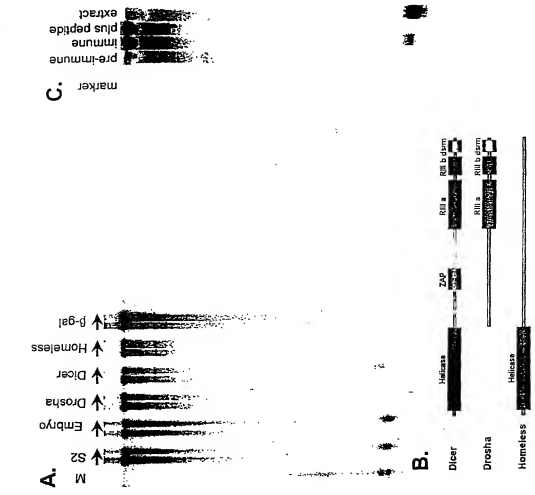

A.
M
S2
Embryo
Drosha
Dicer
Homeless
β-gal
B.
Helicase
ZAP
RIII a
RIII b dsrm
Dicer
Drosha
Homeless
C.
marker
pre-immune
immune
plus peptide
extract

Figure 20

Figure 21

Figure 22

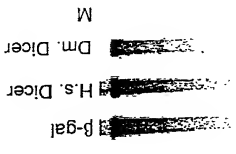


Figure 23

Figure 24

MGKKDKNKKGGQDSAAAPQPQQQQKQQQRQQQPQQLQQPQQLQQPQQLQQPQQQQQQ
QPHQQQQQSSRQQPSTSSGGSRASGFQQGQQQKSQDAEGWTAQKQGKQQVQGWTKQ
GQQGGHQQGRQGQDGGYQQRPPGQQQGGHQQGRQGEGGYQQRPPGQQQGGHQQGRQG
QEGGYQQRPSGQQQGGHQQGRQGQEGGYQQRPPGQQQGGHQQGRQGQEGGYQQRPSGQ
QQGGHQQGRQGQEGGYQQRPSGQQQGGHQQGRQGQEGGYQQRPSGQQQGGHQQGRQGQ
EGGYQQRPPGQQPNQTQSQGQYQSRGPPQQQQRAPLPLPPQPAGSIKRGTIGKPGQVG
INYLDLDLSKMPSVAYHYDVKIMPERPKKFYRQAFEQFRVDQLGGAVLAYDGKASCYS
VDKLPLNSQNFEVTVTDRNGRTLRYTIEIKFTGDSTIDLKSLTTYMNDRIFDKPMRAM
QCVEVVLASPCHNKAIRVGRSFFKWSDPNNRHELDDGYEALVGLYQAFMLGDRPFLNV
DISHKSFPISMPMIEYLERFSLKAKTNNFTNLDYSRRFLEPFLRGINVVYTPPQSFQS
APRVYRVNGLSRAPASSETFEHDGKKVTIASYFHSRNVPLKFPQLHCLNVGSSIKSIL
LPIELCSIEEGQALNRKDGATQVVANMIKYAATSTNVRKRKIMNLLQYFQHNLDPTISR
PGIRIANDFTVVSTRVLSPPQVEYHSKRFTMVKNGSWRMDGMKFLEPKPKAHKCAVLY
CDPRSGRKMNYTQLNDFGNLIISQGKAVNISLDSDVTYRPFTDDERSLDTIIFADLKRS
QHDLAIVIIPQFRISYDTIKQKAELQHGILTQCIKQFTVERKCNNQTIGNILLKINSK
LNGINHKIKDDPRLPMWKNTMYIGADVTHPSPDQREIPSVVGVAASHDPYGASYNMQY
RLQRGALEEIEDMFSITLEHLRVYKEYRNAYPDHIIYYRDGVSDGQFPKIKNEELRCI
KQACDKVGCKPKICCVIVVKRHHTRFFPSGDVTTSNKFNNVDPGTVVDRTIVHPNEMQ
FFMVSHQAIQGTAKPTRYNVIENTGNLDIDLLQQLTYNLCHMFPRCNRSVSYPAPAYIL
AHLVAARGRVVLPGTNNFLDLKKEYAKRTIVPEFMKKNPMYFV

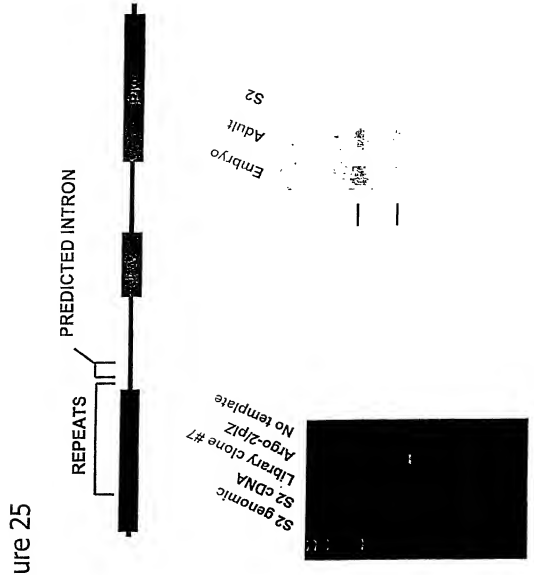

PREDICTED INTRON
REPEATS
S2
Adult
Embryo
S2 genomic
S2 cDNA
Library clone #7
Argo-2plZ
No template

Figure 25

Figure 26

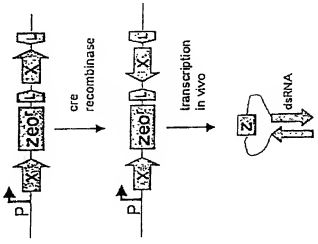

P
loxP
zeo
loxP
X
cre
recombinase
P
loxP
zeo
loxP
X
transcription
in vivo
Z
dsRNA

Figure 27

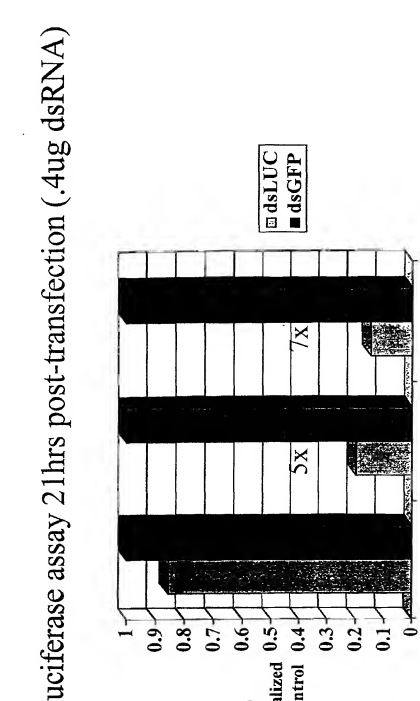

Dual luciferase assay 21hrs post-transfection (.4ug dsRNA)
FF-LUC/
REN-LUC
Values normalized
to dsGFP control
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
5X
7X
dsLUC
dsGFP
Hela
CHO
P19

Figure 28

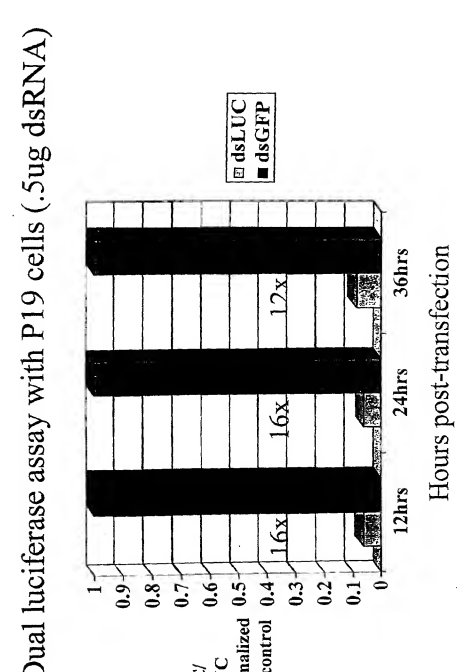

Dual luciferase assay with P19 cells (.5ug dsRNA)
FF-LUC/
REN-LUC
Values normalized
to dsGFP control
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
16x
16x
12x
dsLUC
dsGFP
12hrs
24hrs
36hrs
Hours post-transfection

Figure 29

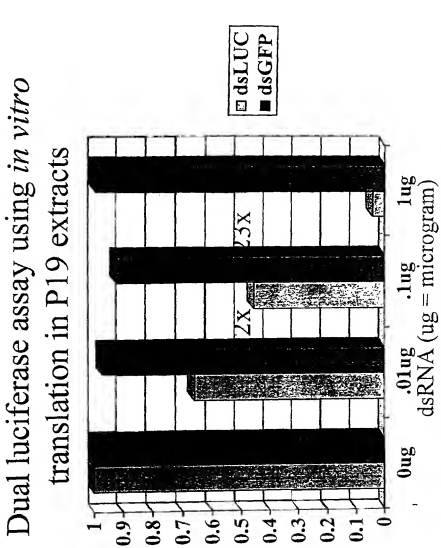

Dual luciferase assay using *in vitro* translation in P19 extracts
FF-LUC/ REN-LUC Values normalized to 0microg control
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
2X
dsLUC
dsGFP
0ug
.01ug
.1ug
1ug
dsRNA (ug = microgram)

Figure 30

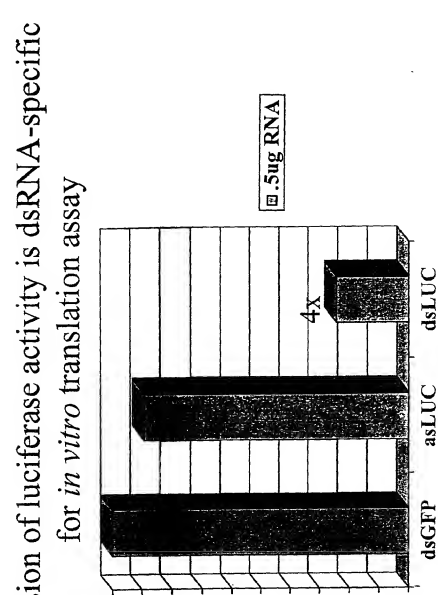

Suppression of luciferase activity is dsRNA-specific for *in vitro* translation assay
FF-LUC/ REN-LUC Values normalized to dsGFP control
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
.5ug RNA
4X
dsGFP
asLUC
dsLUC

Figure 31

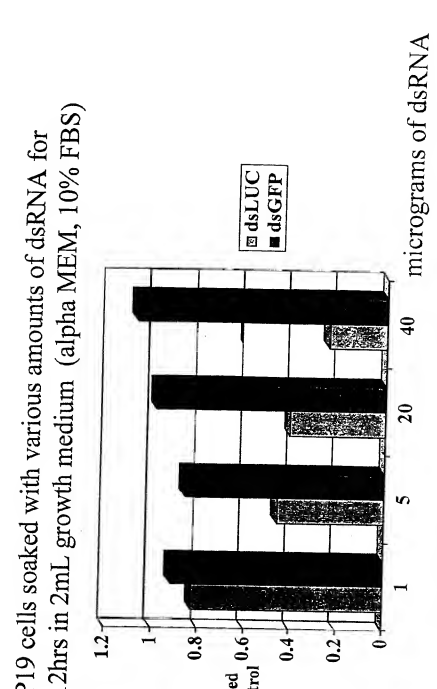
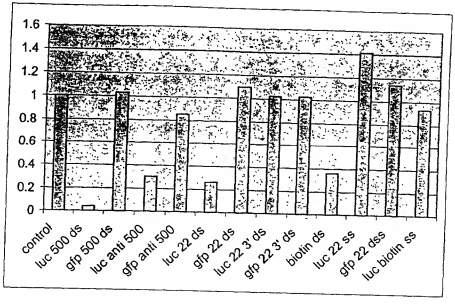

P19 cells soaked with various amounts of dsRNA for 12hrs in 2mL growth medium (alpha MEM, 10% FBS)
FF-LUC/ REN-LUC Values normalized to 0 microg control
1.2
1
0.8
0.6
0.4
0.2
0
1
5
20
40
micrograms of dsRNA
dsLUC
dsGFP

Figure 32

Figure 33


1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
control
luc 500 ds
gfp 500 ds
luc anti 500
gfp anti 500
luc 22 ds
gfp 22 ds
luc 22 3' ds
gfp 22 3' ds
biotin ds
luc 22 ss
gfp 22 dss
luc biotin ss

METHODS AND COMPOSITIONS FOR RNA INTERFERENCE

RELATED APPLICATIONS

[0001] This application is a continuation-in-part of PCT application PCT/US01/08435, filed Mar. 16, 2001, and claims the benefit of U.S. Provisional applications U.S. Ser. No. 60/189,739 filed Mar. 16, 2000 and U.S. Ser. No. 60/243,097 filed Oct. 24, 2000. The specifications of such applications are incorporated by reference herein.

GOVERNMENT SUPPORT

[0002] Work described herein was supported by National Institutes of Health Grant R01-GM62534. The United States Government may have certain rights in the invention.

BACKGROUND OF THE INVENTION

[0003] "RNA interference", "post-transcriptional gene silencing", "quelling"—these different names describe similar effects that result from the overexpression or misexpression of transgenes, or from the deliberate introduction of double-stranded RNA into cells (reviewed in Fire A (1999) *Trends Genet* 15:358-363; Sharp PA (1999) *Genes Dev* 13:139-141; Hunter C (1999) *Curr Biol* 9:R440-R442; Baulcombe DC (1999) *Curr Biol* 9:R599-R601; Vaucheret et al. (1998) *Plant J* 16:651-659). The injection of double-stranded RNA into the nematode *Caenorhabditis elegans*, for example, acts systemically to cause the post-transcriptional depletion of the homologous endogenous RNA (Fire et al. (1998) *Nature* 391: 806-811; and Montgomery et al. (1998) *PNAS* 95:15502-15507). RNA interference, commonly referred to as RNAi, offers a way of specifically and potently inactivating a cloned gene, and is proving a powerful tool for investigating gene function. But the phenomenon is interesting in its own right; the mechanism has been rather mysterious, but recent research—the latest reported by Smardon et al. (2000) *Curr Biol* 10:169-178—is beginning to shed light on the nature and evolution of the biological processes that underlie RNAi.

[0004] RNAi was discovered when researchers attempting to use the antisense RNA approach to inactivate a *C. elegans* gene found that injection of sense-strand RNA was actually as effective as the antisense RNA at inhibiting gene function. Guo et al. (1995) *Cell* 81:611-620. Further investigation revealed that the active agent was modest amounts of double-stranded RNA that contaminate in vitro RNA preparations. Researchers quickly determined the 'rules' and effects of RNAi. Exon sequences are required, whereas introns and promoter sequences, while ineffective, do not appear to compromise RNAi (though there may be gene-specific exceptions to this rule). RNAi acts systemically—injection into one tissue inhibits gene function in cells throughout the animal. The results of a variety of experiments, in *C. elegans* and other organisms, indicate that RNAi acts to destabilize cellular RNA after RNA processing.

[0005] The potency of RNAi inspired Timmons and Fire (1998 *Nature* 395: 854) to do a simple experiment that produced an astonishing result. They fed to nematodes bacteria that had been engineered to express double-stranded RNA corresponding to the *C. elegans* unc-22 gene. Amazingly, these nematodes developed a phenotype similar to that of unc-22 mutants that was dependent on their food source. The ability to conditionally expose large numbers of nematodes to gene-specific double-stranded RNA formed the basis for a very powerful screen to select for RNAi-defective *C. elegans* mutants and then to identify the corresponding genes.

[0006] Double-stranded RNAs (dsRNAs) can provoke gene silencing in numerous in vivo contexts including Drosophila, *Caenorhabditis elegans*, planaria, hydra, trypanosomes, fungi and plants. However, the ability to recapitulate this phenomenon in higher eukaryotes, particularly mammalian cells, has not be accomplished in the art. Nor has the prior art demonstrated that this phenomena can be observe in cultured eukaryotes cells.

SUMMARY OF THE INVENTION

[0007] One aspect of the present invention provides a method for attenuating expression of a target gene in cultured cells, comprising introducing double stranded RNA (dsRNA) into the cells in an amount sufficient to attenuate expression of the target gene, wherein the dsRNA comprises a nucleotide sequence that hybridizes under stringent conditions to a nucleotide sequence of the target gene.

[0008] Another aspect of the present invention provides a method for attenuating expression of a target gene in a mammalian cell, comprising

[0009] (i) activating one or both of a Dicer activity or an Argonaut activity in the cell, and

[0010] (ii) introducing into the cell a double stranded RNA (dsRNA) in an amount sufficient to attenuate expression of the target gene, wherein the dsRNA comprises a nucleotide sequence that hybridizes under stringent conditions to a nucleotide sequence of the target gene.

[0011] In certain embodiments, the cell is suspended in culture; while in other embodiments the cell is in a whole animal, such as a non-human mammal.

[0012] In certain preferred embodiments, the cell is engineered with (i) a recombinant gene encoding a Dicer activity, (ii) a recombinant gene encoding an Argonaut activity, or (iii) both. For instance, the recombinant gene may encode, for a example, a protein which includes an amino acid sequence at least 50 percent identical to SEQ ID No. 2 or 4; or be defined by a coding sequence hybridizes under wash conditions of 2×SSC at 22° C. to SEQ ID No. 1 or 3. In certain embodiments, the recombinant gene may encode, for a example, a protein which includes an amino acid sequence at least 50 percent identical to the Argonaut sequence shown in **FIG. 24**.

[0013] In certain embodiments, rather than use a heterologous expression construct(s), an endogenous Dicer gene or Argonaut gene can be activated, e.g, by gene activation technology, expression of activated transcription factors or other signal transduction protein, which induces expression of the gene, or by treatment with an endogenous factor which upregualtes the level of expression of the protein or inhibits the degradation of the protein.

[0014] In certain preferred embodiments, the target gene is an endogenous gene of the cell. In other embodiments, the

target gene is an heterologous gene relative to the genome of the cell, such as a pathogen gene, e.g., a viral gene.

[0015] In certain embodiments, the cell is treated with an agent that inhibits protein kinase RNA-activated (PKR) apoptosis, such as by treatment with agents which inhibit expression of PKR, cause its destruction, and/or inhibit the kinase activity of PKF.

[0016] In certain preferred embodiments, the cell is a primate cell, such as a human cell.

[0017] In certain preferred embodiments, the length of the dsRNA is at least 20, 21 or 22 nucleotides in length, e.g., corresponding in size to RNA products produced by Dicer-dependent cleavage. In certain embodiments, the dsRNA construct is at least 25, 50, 100, 200, 300 or 400 bases. In certain embodiments, the dsRNA construct is 400-800 bases in length.

[0018] In certain preferred embodiments, expression of the target gene is attenuated by at least 5 fold, and more preferably at least 10, 20 or even 50 fold, e.g., relative to the untreated cell or a cell treated with a dsRNA construct which does not correspond to the target gene.

[0019] Yet another aspect of the present invention provides a method for attenuating expression of a target gene in cultured cells, comprising introducing an expression vector having a "coding sequence" which, when transcribed, produces double stranded RNA (dsRNA) the cell in an amount sufficient to attenuate expression of the target gene, wherein the dsRNA comprises a nucleotide sequence that hybridizes under stringent conditions to a nucleotide sequence of the target gene. An certain embodiments, the vector includes a single coding sequence for the dsRNA which is operably linked to (two) transcriptional regulatory sequences which cause transcription of in both directions (to form complementary transcripts of the coding sequence. In other embodiments, the vector includes two coding sequences which, respectively, give rise to the two complementary sequences which form the dsRNA when annealed. In certain embodiments, the vectors are episomal, e.g., and transfection is transient. In other embodiments, the vectors are chromosomally integrated, e.g., to produce a stably transfected cell line. Preferred vectors for forming such stable cell lines are the described in U.S. Pat. No. 6,025,192 and PCT publication WO/9812339, which are incorporated by reference herein.

[0020] Still another aspect of the present invention provides an assay for identifying nucleic acid sequences responsible for conferring a particular phenotype in a cell, comprising

[0021] (i) constructing a variegated library of nucleic acid sequences from a cell in an orientation relative to a promoter to produce double stranded DNA;

[0022] (ii) introducing the variegated dsRNA library into a culture of target cells, which cells have an activated Dicer activity or Argonaut activity;

[0023] (iii) identifying members of the library which confer a particular phenotype on the cell, and identifying the sequence from a cell which correspond, such as being identical or homologous, to the library member.

[0024] Yet another aspect of the present invention provides a method of conducting a drug discovery business comprising:

[0025] (i) identifying, by the subject assay, a target gene which provides a phenotypically desirable response when inhibited by RNAi;

[0026] (ii) identifying agents by their ability to inhibit expression of the target gene or the activity of an expression product of the target gene;

[0027] (iii) conducting therapeutic profiling of agents identified in step (b), or further analogs thereof, for efficacy and toxicity in animals, and

[0028] (iv) formulating a pharmaceutical preparation including one or more agents identified in step (iii) as having an acceptable therapeutic profile.

[0029] The method may include an additional step of establishing a distribution system for distributing the pharmaceutical preparation for sale, and may optionally include establishing a sales group for marketing the pharmaceutical preparation.

[0030] Another aspect of the present invention provides a method of conducting a target discovery business comprising:

[0031] (i) identifying, by the subject assay, a target gene which provides a phenotypically desirable response when inhibited by RNAi;

[0032] (ii) (optionally) conducting therapeutic profiling of the target gene for efficacy and toxicity in animals; and

[0033] (iii). licensing, to a third party, the rights for further drug development of inhibitors of the target gene.

[0034] Another aspect of the invention provides a method for inhibiting RNAi by inhibiting the expression or activity of an RNAi enzyme. Thus, the subject method may include inhibiting the acitivity of Dicer and/or the 22-mer RNA.

[0035] Still another aspect relates to the a method for altering the specificity of an RNAi by modifying the sequence of the RNA component of the RNAi enzyme.

[0036] Another aspect of the invention relates to purified or semi-purified preparations of the RNAi enzyme or components thereof. In certain embodiments, the preparations are used for identifying compounds, especially small organic molecules, which inhibit or potentiate the RNAi activity. Small molecule inhibitors, for example, can be used to inhibit dsRNA responses in cells which are purposefully being transfected with a virus which produces double stranded RNA.

[0037] The dsRNA construct may comprise one or more strands of polymerized ribonucleotide. It may include modifications to either the phosphate-sugar backbone or the nucleoside. The double-stranded structure may be formed by a single self-complementary RNA strand or two complementary RNA strands. RNA duplex formation may be initiated either inside or outside the cell. The dsRNA construct may be introduced in an amount which allows delivery of at least one copy per cell. Higher doses of double-stranded material may yield more effective inhibition. Inhibition is

sequence-specific in that nucleotide sequences corresponding to the duplex region of the RNA are targeted for genetic inhibition. dsRNA constructs containing a nucleotide sequences identical to a portion of the target gene is preferred for inhibition. RNA sequences with insertions, deletions, and single point mutations relative to the target sequence have also been found to be effective for inhibition. Thus, sequence identity may optimized by alignment algorithms known in the art and calculating the percent difference between the nucleotide sequences. Alternatively, the duplex region of the RNA may be defined functionally as a nucleotide sequence that is capable of hybridizing with a portion of the target gene transcript.

[0038] Yet another aspect of the invention pertains to transgenic non-human mammals which include a transgene encoding a dsRNA construct, preferably which is stably integrated into the genome of cells in which it occurs. The animals can be derived by oocyte microinjection, for example, in which case all of the nucleated cells of the animal will include the transgene, or can be derived using embryonic stem (ES) cells which have been transfected with the transgene, in which case the animal is a chimera and only a portion of its nucleated cells will include the transgene. In certain instances, the sequence-independent dsRNA response, e.g., the PKR response, is also inhibited in those cells including the transgene.

[0039] In still other embodiments, dsRNA itself can be introduced into an ES cell in order to effect gene silencing, and that phenotype will be carried for at least several rounds of division, e.g., into the progeny of that cell.

BRIEF DESCRIPTION OF THE DRAWINGS

[0040] **FIG. 1**: RNAi in S2 cells. a, Drosophila S2 cells were transfected with a plasmid that directs lacZ expression from the copia promoter in combination with dsRNAs corresponding to either human Cu8 or lacZ, or with no dsRNA, as indicated. b, S2 cells were co-transfected with a plasmid that directs expression of a GFP-US9 fusion protein (12) and dsRNAs of either lacZ or cyclin E, as indicated. Upper panels show FACS profiles of the bulk population. Lower panels show FACS profiles from GFP-positive cells. c, Total RNA was extracted from cells transfected with lacZ, cyclin E, fizzy or cyclin A dsRNAs, as indicated. Northern blots were hybridized with sequences not present in the transfected dsRNAs.

[0041] **FIG. 2**: RNAi in vitro. a, Transcripts corresponding to either the first 600 nucleotides of Drosophila cyclin E (E600) or the first 800 nucleotides of lacZ (Z800) were incubated in lysates derived from cells that had been transfected with either lacZ or cyclin E (cycE) dsRNAs, as indicated. Time points were 0, 10, 20, 30, 40 and 60 min for cyclin E and 0, 10, 20, 30 and 60 min for lacZ. b, Transcripts were incubated in an extract of S2 cells that had been transfected with cyclin E dsRNA (cross-hatched box, below). Transcripts corresponded to the first 800 nucleotides of lacZ or the first 600, 300, 220 or 100 nucleotides of cyclin E, as indicated. Eout is a transcript derived from the portion of the cyclin E cDNA not contained within the transfected dsRNA. E-ds is identical to the dsRNA that had been transfected into S2 cells. Time points were 0 and 30 min. c, Synthetic transcripts complementary to the complete cyclin E cDNA (Eas) or the final 600 nucleotides (Eas600) or 300 nucleotides (Eas300) were incubated in extract for 0 or 30 min.

[0042] **FIG. 3**: Substrate requirements of the RISC. Extracts were prepared from cells transfected with cyclin E dsRNA. Aliquots were incubated for 30 min at 30° C. before the addition of either the cyclin E (E600) or lacZ (Z800) substrate. Individual 20-μl aliquots, as indicated, were pre-incubated with 1 mM $CaCl_2$ and 5 mM EGTA, 1 mM $CaCl_2$, 5 mM EGTA and 60 U of micrococcal nuclease, 1 mM $CaCl_2$ and 60 U of micrococcal nuclease or 10 U of DNase I (Promega) and 5 mM EGTA. After the 30-min pre-incubation, EGTA was added to those samples that lacked it. Yeast tRNA (1 μg) was added to all samples. Time points were at 0 and 30 min.

[0043] **FIG. 4**: The RISC contains a potential guide RNA. a, Northern blots of RNA from either a crude lysate or the S100 fraction (containing the soluble nuclease activity, see Methods) were hybridized to a riboprobe derived from the sense strand of the cyclin E mRNA. b, Soluble cyclin-E-specific nuclease activity was fractionated as described in Methods. Fractions from the anion-exchange resin were incubated with the lacZ, control substrate (upper panel) or the cyclin E substrate (centre panel). Lower panel, RNA from each fraction was analysed by northern blotting with a uniformly labelled transcript derived from sense strand of the cyclin E cDNA. DNA oligonucleotides were used as size markers.

[0044] **FIG. 5**: Generation of 22 mers and degradation of mRNA are carried out by distinct enzymatic complexes. A. Extracts prepared either from 0-12 hour Drosophila embryos or Drosophila S2 cells (see Methods) were incubated 0, 15, 30, or 60 minutes (left to right) with a uniformly-labeled double-stranded RNA corresponding to the first 500 nucleotides of the Drosophila cyclin E coding region. M indicates a marker prepared by in vitro transcription of a synthetic template. The template was designed to yield a 22 nucleotide transcript. The doublet most probably results from improper initiation at the +1 position. B. Whole-cell extracts were prepared from S2 cells that had been transfected with a dsRNA corresponding to the first 500 nt. of the luciferase coding region. S10 extracts were spun at 30,000xg for 20 minutes which represents our standard RISC extract[6]. S100 extracts were prepared by further centrifugation of S10 extracts for 60 minutes at 100,000xg. Assays for mRNA degradation were carried out as described previously[6] for 0, 30 or 60 minutes (left to right in each set) with either a single-stranded luciferase mRNA or a single-stranded cyclin E mRNA, as indicated. C. S10 or S100 extracts were incubated with cyclin E dsRNAs for 0, 60 or 120 minutes (L to R).

[0045] **FIG. 6**: Production of 22 mers by recombinant CG4792/Dicer. A. Drosophila S2 cells were transfected with plasmids that direct the expression of T7-epitope tagged versions of Drosha, CG4792/Dicer-1 and Homeless. Tagged proteins were purified from cell lysates by immunoprecipitation and were incubated with cyclin E dsRNA. For comparison, reactions were also performed in Drosophila embryo and S2 cell extracts. As a negative control, immunoprecipitates were prepared from cells transfected with a β-galactosidase expression vector. Pairs of lanes show reactions performed for 0 or 60 minutes. The synthetic marker (M) is as described in the legend to **FIG. 1**. B. Diagrammatic representations of the domain structures of CG4792/Dicer-1, Drosha and Homeless are shown. C. Immunoprecipitates were prepared from detergent lysates of S2 cells using an

antiserum raised against the C-terminal 8 amino acids of Drosophila Dicer-1 (CG4792). As controls, similar preparations were made with a pre-immune serum and with an immune serum that had been pre-incubated with an excess of antigenic peptide. Cleavage reactions in which each of these precipitates was incubated with an ~500 nt. fragment of Drosophila cyclin E are shown. For comparsion, an incubation of the substrate in Drosophila embryo extract was electrophoresed in parallel. D. Dicer immunoprecipitates were incubated with dsRNA substrates in the presence or absence of ATP. For comparison, the same substrate was incubated with S2 extracts that either contained added ATP or that were depleted of ATP using glucose and hexokinase (see methods). E. Drosophila S2 cells were transfected with uniformly, 32P-labelled dsRNA corresponding to the first 500 nt. of GFP. RISC complex was affinity purified using a histidine-tagged version of D.m. Ago-2, a recently identified component of the RISC complex (Hammond et al., in prep). RISC was isolated either under conditions in which it remains ribosome associated (ls, low salt) or under conditions that extract it from the ribosome in a soluble form (hs, high salt)[6]. For comparison, the spectrum of labelled RNAs in the total lysate is shown. F. Guide RNAs produced by incubation of dsRNA with a Dicer immunoprecipitate are compared to guide RNAs present in a affinity-purified RISC complex. These precisely comigrate on a gel that has single-nucleotide resolution. The lane labelled control is an affinity selection for RISC from cell that had been transfected with labeled dsRNA but not with the epitope-tagged D.m. Ago-2.

[0046] FIG. 7: Dicer participates in RNAi. A. Drosophila S2 cells were transfected with dsRNAs corresponding to the two Drosophila Dicers (CG4792 and CG6493) or with a control dsRNA corresponding to murine caspase 9. Cytoplasmic extracts of these cells were tested for Dicer activity. Transfection with Dicer dsRNA reduced activity in lysates by 7.4-fold. B. The Dicer-1 antiserum (CG4792) was used to prepare immunoprecipitates from S2 cells that had been treated as described above. Dicer dsRNA reduced the activity of Dicer-1 in this assay by 6.2-fold. C. Cells that had been transfected two days previously with either mouse caspase 9 dsRNA or with Dicer dsRNA were cotransfected with a GFP expression plasmid and either control, luciferase dsRNA or GFP dsRNA. Three independent experiments were quantified by FACS. A comparison of the relative percentage of GFP-positive cells is shown for control (GFP plasmid plus luciferase dsRNA) or silenced (GFP plamsid plus GFP dsRNA) populations in cells that had previously been transfected with either control (caspase 9) or Dicer dsRNAs.

[0047] FIG. 8: Dicer is an evolutionarily conserved ribonuclease. A. A model for production of 22 mers by Dicer. Based upon the proposed mecbanism of action of Ribonuclease III, we propose that Dicer acts on its substrate as a dimer. The positioning of the two ribonuclease domains (RIIIa and RIIIb) within the enzyme would thus determine the size of the cleavage product. An equally plausible alternative model could be derived in which the RIIIa and RIIIb domains of each Dicer enzyme would cleave in concert at a single position. In this model, the size of the cleavage product would be determined by interaction between two neighboring Dicer enzymes. B. Comparison of the domain structures of potential Dicer homologs in various organisms (Drosophila—CG4792, CG6493, *C. elegans*—K12H4.8, Arabidopsis—CARPEL FACTORY[24], T25K16.4, AC012328_1, human Helicase-MOI[25] and *S. pombe*—YC9A_SCHPO). The ZAP domains were identified both by analysis of individual sequences with Pfam[27] and by Psi-blast[28] searches. The ZAP domain in the putative *S. pombe* Dicer is not detected by PFAM but is identified by Psi-Blast and is thus shown in a different color. For comparison, a domain structure of the RDE1/QDE2/ARGONAUTE family is shown. It should be noted that the ZAP domains are more similar within each of the Dicer and ARGONAUTE families than they are between the two groups. C. An alignment of the ZAP domains in selected Dicer and Argonaute family members is shown. The alignment was produced using ClustalW.

[0048] FIG. 9: Purification strategy for RISC. (second step in RNAi model).

[0049] FIG. 10: Fractionation of RISC activity over sizing column. Activity fractionates as 500 KD complex. Also, antibody to dm argonaute 2 cofractionates with activity.

[0050] FIGS. **11-13**: Fractionation of RISC over monoS, monoQ, Hydroxyapatite columns. Dm argonaute 2 protein also cofactionates.

[0051] FIG. 14: Alignment of dm argonaute 2 with other family members.

[0052] FIG. 15: Confirmation of dm argonaute 2. S2 cells were transfected with labeled dsRNA and His tagged argonaute. Argonaute was isolated on nickel agarose and RNA component was identified on 15% acrylamide gel.

[0053] FIG. 16: S2 cell and embryo extracts were assayed for 22 mer generating activity.

[0054] FIG. 17: RISC can be separated from 22 mer generating activity (dicer). Spinning extracts (S100) can clear RISC activity from supernatant (left panel) however, S100 spins still contain dicer activity (right panel).

[0055] FIG. 18: Dicer is specific for dsRNA and prefers longer substrates.

[0056] FIG. 19: Dicer was fractionated over several columns.

[0057] FIG. 20: Identification of dicer as enzyme which can process dsRNA into 22 mers. Various RNaseIII family members were expressed with n terminal tags, immunoprecipitated, and assayed for 22 mer generating activity (left panel). In right panel, antibodies to dicer could also precipitate 22 mer generating activity.

[0058] FIG. 21: Dicer requires ATP.

[0059] FIG. 22: Dicer produces RNAs that are the same size as RNAs present in RISC.

[0060] FIG. 23: Human dicer homolog wben expressed and immunoprecipitated has 22 mer generating activity.

[0061] FIG. 24: Sequence of dm argonaute 2. Peptides identified by microsequencing are shown in underline.

[0062] FIG. 25: Molecular charaterization of dm argonaute 2. The presence of an intron in coding sequence was determined by northern blotting using intron probe. This results in a different 5' reading frame that that published genome sequence. Number of polyglutaine repeats was determined by genomic PCR.

[0063] **FIG. 26**: Dicer activity can be created in human cells by expression of human dicer gene. Host cell was 293. Crude extracts had dicer activity, while activity was absent from untransfected cells. Activity is not dissimilar to that seen in drosophila embryo extracts.

[0064] **FIG. 27**: An ~500 nt. fragment of the gene that is to be silenced (X) is inserted into the modified vector as a stable direct repeat using standard cloning procedures. Treatment with commercially available cre recombinase reverses sequences within the loxP sites (L) to create an inverted repeat. This can be stably maintained and amplified in an sbc mutant bacterial strain (DL759). Transcription in vivo from the promoter of choice (P) yields a hairpin RNA that causes silencing. A zeocin resistance marker is included to insure maintenance of the direct and inverted repeat structures; however this is non-essential in vivo and could be removed by pre-mRNA splicing if desired. Smith, N. A. et al. Total silencing by intron-spliced hairpin RNAs. *Nature* 407, 319-20 (2000).

[0065] **FIG. 28**: Hela, Chinese hamster ovary, and P19 (pluripotent, mouse embryonic carcinoma) cell lines transfected with plasmids expressing Photinus pyralis (firefly) Renilla reniformis (sea pansy) luciferases and with dsRNA 500 mers (400 ng), either homologous to firefly luciferase mRNA (dsLUC) or non-homologous (dsGFP). Dual luciferase assays were carried out using an Analytical Scientific Instruments model 3010 Luminometer. In this assay Renilla luciferase serves as an internal control for dsRNA-specific suppression of firefly luciferase activity. These data demonstrate that 500 mer dsRNA can specifically suppress cognate gene expression in vivo.

[0066] **FIG. 29**: P19 (a pluripotent, mouse embryonic cell line) cells transfected with plasmids expressing Photinus pyralis (firefly) Renilla reniformis (sea pansy) luciferases and with dsRNA 500 mers (500ng), either homologous to firefly luciferase mRNA (dsLUC) or non-homologous (dsGFP). Dual luciferase assays were carried out using an Analytical Scientific Instruments model 3010 Luminometer. In this assay Renilla luciferase serves as an internal control for dsRNA-specific suppression of firefly luciferase activity. These data further demonstrate that 500 mer dsRNA can specifically suppress cognate gene expression in vivo and that the effect is stable over time .

[0067] **FIG. 30**: S10 fractions from P19 cell lysates were used for in vitro translations of mRNA coding for Photinus pyralis (firefly) Renilla reniformis (sea pansy) luciferases. Translation reactions were programmed with various amounts of dsRNA 500 mers, either homologous to firefly luciferase mRNA (dsLUC) or non-homologous (dsGFP). Reactions were carried out at 30 degrees for 1 hour, after which dual luciferase assays were carried out using an Analytical Scientific Instruments model 3010 Luminometer. In this assay Renilla luciferase serves as an internal control for dsRNA-specific suppression of firefly luciferase activity. These data demonstrate that 500 mer dsRNA can specifically suppress cognate gene expression in vitro in a manner consistent with post-transcriptional gene silencing. Antisense firefly RNA did not differ significantly from dsGFP control (approximately 10%) (data not shown).

[0068] **FIG. 31**: S10 fractions from P19 cell lysates were used for in vitro translations of mRNA coding for Photinus pyralis (firefly) Renilla reniformis (sea pansy) luciferases. Translation reactions were programmed with dsRNA or asRNA 500 mers, either complementary to firefly luciferase mRNA (asLUC and dsLUC) or non-complementary (dsGFP). Reactions were carried out at 30 degrees for 1 hour, after a 30 min preincubation with dsRNA or asRNA. Dual luciferase assays were carried out using an Analytical Scientific Instruments model 3010 Luminometer. In this assay Renilla luciferase serves as an internal control for dsRNA-specific suppression of firefly luciferase activity. These data demonstrate that 500 mer double-stranded RNA (dsRNA) but not anti-sense RNA (asRNA) suppresses cognate gene expression in vitro in a manner consistent with post-transcriptional gene silencing.

[0069] **FIG. 32**: P19 cells were grown in 6-well tissue culture plates to approximately 60% confluence. Various amounts of dsRNA, either homologous to firefly luciferase mRNA (dsLUC) or non-homologous (dsGFP), were added to each well and incubated for 12 hrs under normal tissue culture conditions. Cells were then transfected with plasmids expressing Photinus pyralis (firefly) Renilla reniformis (sea pansy) luciferases and with dsRNA 500 mers (500 ng). Dual luciferase assays were carried out 12 hrs post-transfection using an Analytical Scientific Instruments model 3010 Luminometer. In this assay Renilla luciferase serves as an internal control for dsRNA-specific suppression of firefly luciferase activity. These data show that 500 mer dsRNA can specifically suppress cognate gene expression in vivo without transfection under normal tissue culture conditions.

[0070] **FIG. 33**: Is a graph illustrating the relative rate of expression luciferase in cells which are treated with various antisense and dsRNA constructs.

DETAILED DESCRIPTION OF THE CERTAIN PREFERRED EMBODIMENTS

I. Overview

[0071] The present invention provides methods for attenuating gene expression in a cell using gene-targeted double stranded RNA (dsRNA). The dsRNA contains a nucleotide sequence that hybridizes under physiologic conditions of the cell to the nucleotide sequence of at least a portion of the gene to be inhibited (the "target" gene).

[0072] A significant aspect to certain embodiments of the present invention relates to the demonstration in the present application that RNAi can in fact be accomplished in cultured cells, rather than whole organisms as described in the art.

[0073] Another salient feature of the present invention concerns the ability to carry out RNAi in higher eukaryotes, particularly in non-oocytic cells of mammals, e.g., cells from adult mammals as an example.

[0074] As described in further detail below, the present invention(s) are based on the discovery that the RNAi phenomenum is mediated by a set of enzyme activities, including an essential RNA component, that are evolutionarily conserved in eukaryotes ranging from plants to mammals.

[0075] One enzyme contains an essential RNA component. After partial purification, a multi-component nuclease (herein "RISC nuclease") co-fractionates with a discrete, 22-nucleotide RNA species which may confer specificity to

the nuclease through homology to the substrate mRNAs. The short RNA molecules are generated by a processing reaction from the longer input dsRNA. Without wishing to be bound by any particular theory, these 22 mer guide RNAs may serve as guide sequences that instruct the RISC nuclease to destroy specific mRNAs corresponding to the dsRNA sequences.

[0076] As illustrated in FIG. 33, double stranded forms of the 22-mer guide RNA can be sufficient in length to induce sequence-dependent dsRNA inhibition of gene expression. In the illustrated example, dsRNA contructs are administered to cells having a recombinant luciferase reporter gene. The control cell, e.g., no exogeneously added RNA, the level of expression of the luciferase reporter is normalized to be the value of "1". As illustrated, both long (500-mer) and short (22-mer) dsRNA constructs complementary to the luciferase gene could inhibit expression of that gene product relative to the control cell. On the other hand, similarly sized dsRNA complementary to the coding sequence for another protein, green fluorescence protein (GFP), did not significantly effect the expression of luciferase—indicating that the inhibitory phenomena was in each case sequence-dependent. Likewise, single stranded 22-mers of luciferase did not inhibit expression of that gene—indicating that the inhibitory phenomena is double stranded-dependent.

[0077] The appended examples also identify an enzyme, Dicer, that can produce the putative guide RNAs. Dicer is a member of the RNAse III family of nucleases that specifically cleave dsRNA and is evolutionarily conserved in worms, flies, plants, fungi and, as described herein, mammals. The enzyme has a distinctive structure which includes a helicase domain and dual RNAse III motifs. Dicer also contains a region of homology to the RDE1/QDE2/ARGONAUTE family, which have been genetically linked to RNAi in lower eukaryotes. Indeed, activation of, or overexpression of Dicer may be sufficient in many cases to permit RNA interference in otherwise non-receptive cells, such as cultured eukaryotic cells, or mammalian (non-oocytic) cells in culture or in whole organisms.

[0078] In certain embodiments, the cells can be treated with an agent(s) that inhibits the general double-stranded RNA response(s) by the host cells, such as may give rise to sequence-independent apoptosis. For instance, the cells can be treated with agents that inhibit the dsRNA-dependent protein kinase known as PKR (protein kinase RNA-activated). Double stranded RNAs in mammalian cells typically activate protein kinase PKR and leads to apoptosis. The mechanism of action of PKR includes phosphorylation and inactivation eIF2a (Fire (1999) *Trends Genet* 15:358). It has also been reported that induction of NF-κB by PKR is involved in apoptosis commitment and this process is mediated through activation of the IKK complex. This sequence-independent response may reflect a form of primitive immune response, since the presence of dsRNA is a common feature of many viral lifecycles.

[0079] As described herein, Applicants have demonstrated that the PKR response can be overcome in favor of the sequence-specific RNAi response. However, in certain instances, it can be desirable to treat the cells with agents which inhibit expression of PKR, cause its destruction, and/or inhibit the kinase activity of PKF are specifically contemplated for use in the present method. Likewise, overexpression of or agents which ectopic activate IF2αa. can be used. Other agents which can be used to suppress the PKR response include inhibitors of IKK phosphorylation of IκB, inhibitors of IκB ubiquitination, inhibitors of IκB degradation, inhibitors of NF-κB nuclear translocation, and inhibitors of NF-κB interaction with κB response elements.

[0080] Other inhibitors of sequence-independent dsRNA response in cells include the gene product of the vaccinia virus E3L. The E3L gene product contains two distinct domains. A conserved carboxy-terminal domain has been shown to bind double-stranded RNA (dsRNA) and inhibit the antiviral dsRNA response by cells. Expression of at least that portion of the E3L gene in the host cell, or the use of polypeptide or peptidomimetics thereof, can be used to suppress the general dsRNA response. Caspase inhibitors sensitized cells to killing by double-stranded RNA. Accordingly, ectopic expression or activated of caspases in the host cell can be used to suppress the general dsRNA response.

[0081] In other embodiments, the subject method is carried out in cells which have little or no general response to double stranded RNA, e.g., have no PKR-dependent dsRNA response, at least under the culture conditions. As illustrated in FIGS. 28-32, CHO and P19 cells can be used without having to inhibit PKR or other general dsRNA responses.

[0082] Thus, the present invention provides a process and compositions for inhibiting expression of a target gene in a cell, expecially a mammalian cell. In certain embodiments, the process comprises introduction of RNA (the "dsRNA construct") with partial or fully double-stranded character into the cell or into the extracellular environment. Inhibition is specific in that a nucleotide sequence from a portion of the target gene is chosen to produce the dsRNA construct. In preferred embodiments, the method utilizes a cell in which Dicer and/or Argonaute activities are recombinantly expressed or otherwise ectopically activated. This process can be (1) effective in attenuating gene expression, (2) specific to the targeted gene, and (3) general in allowing inhibition of many different types of target gene.

II. Definitions

[0083] For convenience, certain terms employed in the specification, examples, and appended claims are collected here.

[0084] As used herein, the term "vector" refers to a nucleic acid molecule capable of transporting another nucleic acid to that it has been linked. One type of vector is a genomic integrated vector, or "integrated vector", which can become integrated into the chromsomal DNA of the host cell. Another type of vector is an episomal vector, i.e., a nucleic acid capable of extra-chromosomal replication. Vectors capable of directing the expression of genes to that they are operatively linked are referred to herein as "expression vectors". In the present specification, "plasmid" and "vector" are used interchangeably unless otherwise clear from the context.

[0085] As used herein, the term "nucleic acid" refers to polynucleotides such as deoxyribonucleic acid (DNA), and, where appropriate, ribonucleic acid (RNA). The term should also be understood to include, as applicable to the embodiment being described, single-stranded (such as sense or antisense) and double-stranded polynucleotides.

[0086] As used herein, the term "gene" or "recombinant gene" refers to a nucleic acid comprising an open reading frame encoding a polypeptide of the present invention, including both exon and (optionally) intron sequences. A "recombinant gene" refers to nucleic acid encoding such regulatory polypeptides, that may optionally include intron sequences that are derived from chromosomal DNA. The term "intron" refers to a DNA sequence present in a given gene that is not translated into protein and is generally found between exons. As used herein, the term "transfection" means the introduction of a nucleic acid, e.g., an expression vector, into a recipient cell by nucleic acid-mediated gene transfer.

[0087] A "protein coding sequence" or a sequence that "encodes" a particular polypeptide or peptide, is a nucleic acid sequence that is transcribed (in the case of DNA) and is translated (in the case of mRNA) into a polypeptide in vitro or in vivo when placed under the control of appropriate regulatory sequences. The boundaries of the coding sequence are determined by a start codon at the 5' (amino) terminus and a translation stop codon at the 3' (carboxy) terminus. A coding sequence can include, but is not limited to, cDNA from procaryotic or eukaryotic mRNA, genomic DNA sequences from procaryotic or eukaryotic DNA, and even synthetic DNA sequences. A transcription termination sequence will usually be located 3' to the coding sequence.

[0088] Likewise, "encodes", unless evident from its context, will be meant to include DNA sequences that encode a polypeptide, as the term is typically used, as well as DNA sequences that are transcribed into inhibitory antisense molecules.

[0089] The term "loss-of-function", as it refers to genes inhibited by the subject RNAi method, refers a diminishment in the level of expression of a gene when compared to the level in the absense of dsRNA constructs.

[0090] The term "expression" with respect to a gene sequence refers to transcription of the gene and, as appropriate, translation of the resulting mRNA transcript to a protein. Thus, as will be clear from the context, expression of a protein coding sequence results from transcription and translation of the coding sequence.

[0091] "Cells,""host cells" or "recombinant host cells" are terms used interchangeably herein. It is understood that such terms refer not only to the particular subject cell but to the progeny or potential progeny of such a cell. Because certain modifications may occur in succeeding generations due to either mutation or environmental influences, such progeny may not, in fact, be identical to the parent cell, but are still included within the scope of the term as used herein.

[0092] The term "cultured cells" refers to cells suspended in culture, e.g., dispersed in culture or in the form tissue. It does not, however, include oocytes or whole embryos (including blastocysts and the like) which may be provided in culture. In certain embodiments, the cultured cells are adults cells, e.g., non-embryonic.

[0093] By "recombinant virus" is meant a virus that has been genetically altered, e.g., by the addition or insertion of a heterologous nucleic acid construct into the particle.

[0094] As used herein, the terms "transduction" and "transfection" are art recognized and mean the introduction of a nucleic acid, e.g., an expression vector, into a recipient cell by nucleic acid-mediated gene transfer. "Transformation", as used herein, refers to a process in which a cell's genotype is changed as a result of the cellular uptake of exogenous DNA or RNA, and, for example, the transformed cell expresses a dsRNA contruct.

[0095] "Transient transfection" refers to cases where exogenous DNA does not integrate into the genome of a transfected cell, e.g., where episomal DNA is transcribed into mRNA and translated into protein.

[0096] A cell has been "stably transfected" with a nucleic acid construct when the nucleic acid construct is capable of being inherited by daughter cells.

[0097] As used herein, a "reporter gene construct" is a nucleic acid that includes a "reporter gene" operatively linked to at least one transcriptional regulatory sequence. Transcription of the reporter gene is controlled by these sequences to which they are linked. The activity of at least one or more of these control sequences can be directly or indirectly regulated by the target receptor protein. Exemplary transcriptional control sequences are promoter sequences. A reporter gene is meant to include a promoter-reporter gene construct that is heterologously expressed in a cell.

[0098] As used herein, "transformed cells" refers to cells that have spontaneously converted to a state of unrestrained growth, i.e., they have acquired the ability to grow through an indefinite number of divisions in culture. Transformed cells may be characterized by such terms as neoplastic, anaplastic and/or hyperplastic, with respect to their loss of growth control. For purposes of this invention, the terms "transformed phenotype of malignant mammalian cells" and "transformed phenotype " are intended to encompass, but not be limited to, any of the following phenotypic traits associated with cellular transformation of mammalian cells: immortalization, morphological or growth transformation, and tumorigenicity, as detected by prolonged growth in cell culture, growth in semi-solid media, or tumorigenic growth in immuno-incompetent or syngeneic animals.

[0099] As used herein, "proliferating" and "proliferation" refer to cells undergoing mitosis.

[0100] As used herein, "immortalized cells" refers to cells that have been altered via chemical, genetic, and/or recombinant means such that the cells have the ability to grow through an indefinite number of divisions in culture.

[0101] The "growth state" of a cell refers to the rate of proliferation of the cell and the state of differentiation of the cell.

III. Exemplary Embodiments of Isolation Method

[0102] One aspect of the invention provides a method for potentiating RNAi by induction or ectopic activation of an RNAi enzyme in a cell (in vivo or in vitro) or cell-free mixtures. In preferred embodiments, the RNAi activity is activated or added to a mammalian cell, e.g., a human cell, which cell may be provided in vitro or as part of a whole organism. In other embodiments, the subject method is carried out using eukaryotic cells generally (except for oocytes) in culture. For instance, the Dicer enzyme may be activated by virtue of being recombinantly expressed or it

may be activated by use of an agent which (i) induces expression of the endogenous gene, (ii) stabilizes the protein from degradation, and/or (iii) allosterically modies the enzyme to increase its activity (by altering its Kcat, Km or both).

[0103] A. Dicer and Argonaut Activities

[0104] In certain embodiment, at least one of the activated RNAi enzymes is Dicer, or a homolog thereof. In certain preferred embodiments, tbe present method provides for ectopic activation of Dicer. As used herein, the term "Dicer" refers to a protein which (a) mediates an RNAi response and (b) has an amino acid sequence at least 50 percent identical, and more preferablty at least 75, 85, 90 or 95 percent identical to SEQ ID No. 2 or 4, and/or which can be encoded by a nucleic acid which hybridizes under wash conditions of 2×SSC at 22° C., and more preferably 0.2×SSC at 65° C., to a nucleotide represented by SEQ ID No. 1 or 3. Accordingly, the method may comprise introducing a dsRNA contruct into a cell in which Dicer has been recombinantly expressed or otherwise ectopically activated.

[0105] In certain embodiment, at least one of the activated RNAi enzymes is Argonaut, or a homolog thereof. In certain preferred embodiments, the present method provides for ectopic activation of Argonaut. As used herein, the term "Argonaut" refers to a protein which (a) mediates an RNAi response and (b) has an amino acid sequence at least 50 percent identical, and more preferablty at least 75, 85, 90 or 95 percent identical to the amino acid sequence shown in FIG. 24. Accordingly, the method may comprise introducing a dsRNA contruct into a cell in which Argonaut has been recombinantly expressed or otherwise ectopically activated.

[0106] This invention also provides expression vectors containing a nucleic acid encoding a Dicer or Argonaut polypeptides, operably linked to at least one transcriptional regulatory sequence. Operably linked is intended to mean that the nucleotide sequence is linked to a regulatory sequence in a manner which allows expression of the nucleotide sequence. Regulatory sequences are art-recognized and are selected to direct expression of the subject Dicer or Argonaut proteins. Accordingly, the term transcriptional regulatory sequence includes promoters, enhancers and other expression control elements. Such regulatory sequences are described in Goeddel; Gene Expression Technology: Methods in Enzymology 185, Academic Press, San Diego, Calif. (1990). For instance, any of a wide variety of expression control sequences, sequences that control the expression of a DNA sequence when operatively linked to it, may be used in these vectors to express DNA sequences encoding Dicer or Argonaut polypeptides of this invention. Such useful expression control sequences, include, for example, a viral LTR, such as the LTR of the Moloney murine leukemia virus, the early and late promoters of SV40, adenovirus or cytomegalovirus immediate early promoter, tbe lac system, the trp system, the TAC or TRC system, T7 promoter whose expression is directed by T7 RNA polymerase, the major operator and promoter regions of phage λ, the control regions for fd coat protein, the promoter for 3-phosphoglycerate kinase or other glycolytic enzymes, the promoters of acid phospbatase, e.g., Pho5, the promoters of the yeast α-mating factors, the polybedron promoter of the baculovirus system and other sequences known to control the expression of genes of prokaryotic or eukaryotic cells or their viruses, and various combinations thereof. It sbould be understood that the design of the expression vector may depend on such factors as the choice of the host cell to be transformed and/or the type of protein desired to be expressed.

[0107] Moreover, the vector's copy number, the ability to control that copy number and the expression of any other proteins encoded by the vector, such as antibiotic markers, should also be considered.

[0108] The recombinant Dicer or Argonaut genes can be produced by ligating nucleic acid encoding a Dicer or Argonaut polypeptide into a vector suitable for expression in either prokaryotic cells, eukaryotic cells, or both. Expression vectors for production of recombinant forms of the subject Dicer or Argonaut polypeptides include plasmids and other vectors. For instance, suitable vectors for the expression of a Dicer or Argonaut polypeptide include plasmids of the types: pBR322-derived plasmids, pEMBL-derived plasmids, pEX-derived plasmids, pBTac-derived plasmids and pUC-derived plasmids for expression in prokaryotic cells, such as *E. coli*.

[0109] A number of vectors exist for the expression of recombinant proteins in yeast. For instance, YEP24, YIP5, YEP51, YEP52, pYES2, and YRP17 are cloning and expression vehicles useful in the introduction of genetic constructs into *S. cerevisiae* (see, for example, Broach et al. (1983) in Experimental Manipulation of Gene Expression, ed. M. Inouye Academic Press, p. 83, incorporated by reference herein). These vectors can replicate in *E. coli* due the presence of the pBR322 ori, and in *S. cerevisiae* due to the replication determinant of the yeast 2 micron plasmid. In addition, drug resistance markers such as ampicillin can be used. In an illustrative embodiment, a Dicer or Argonaut polypeptide is produced recombinantly utilizing an expression vector generated by sub-cloning the coding sequence of a Dicer or Argonaut gene.

[0110] The preferred mammalian expression vectors contain both prokaryotic sequences, to facilitate the propagation of the vector in bacteria, and one or more eukaryotic transcription units that are expressed in eukaryotic cells. The pcDNAI/amp, pcDNAI/neo, pRc/CMV, pSV2gpt, pSV2neo, pSV2-dhfr, pTk2, pRSVneo, pMSG, pSVT7, pko-neo and pHyg derived vectors are examples of mammalian expression vectors suitable for transfection of eukaryotic cells. Some of these vectors are modified with sequences from bacterial plasmids, such as pBR322, to facilitate replication and drug resistance selection in both prokaryotic and eukaryotic cells. Alternatively, derivatives of viruses such as the bovine papillomavirus (BPV-1), or Epstein-Barr virus (pHEBo, pREP-derived and p205) can be used for transient expression of proteins in eukaryotic cells. The various methods employed in the preparation of the plasmids and transformation of host organisms are well known in the art. For other suitable expression systems for both prokaryotic and eukaryotic cells, as well as general recombinant procedures, see Molecular Cloning A Laboratory Manual, 2nd Ed., ed. by Sambrook, Fritsch and Maniatis (Cold Spring Harbor Laboratory Press: 1989) Chapters 16 and 17.

[0111] In yet another embodiment, the subject invention provides a "gene activation" construct which, by homologous recombination with a genomic DNA, alters the transcriptional regulatory sequences of an endogenous Dicer or

Argonaut gene. For instance, the gene activation construct can replace the endogenous promoter of a Dicer or Argonaut gene with a heterologous promoter, e.g., one which causes constitutive expression of the Dicer or Argonaut gene or which causes inducible expression of the gene under conditions different from the normal expression pattern of Dicer or Argonaut. A variety of different formats for the gene activation constructs are available. See, for example, the Transkaryotic Therapies, Inc PCT publications WO93/09222, WO95/31560, WO96/29411, WO95/31560 and WO94/12650.

[0112] In preferred embodiments, the nucleotide sequence used as the gene activation construct can be comprised of (1) DNA from some portion of the endogenous Dicer or Argonaut gene (exon sequence, intron sequence, promoter sequences, etc.) which direct recombination and (2) heterologous transcriptional regulatory sequence(s) which is to be operably linked to the coding sequence for the genomic Dicer or Argonaut gene upon recombination of the gene activation construct. For use in generating cultures of Dicer or Argonaut producing cells, the construct may further include a reporter gene to detect the presence of the knockout construct in the cell.

[0113] The gene activation construct is inserted into a cell, and integrates with the genomic DNA of the cell in such a position so as to provide the heterologous regulatory sequences in operative association with the native Dicer or Argonaut gene. Such insertion occurs by homologous recombination, i.e., recombination regions of the activation construct that are homologous to the endogenous Dicer or Argonaut gene sequence hybridize to the genomic DNA and recombine with the genomic sequences so that the construct is incorporated into the corresponding position of the genomic DNA.

[0114] The terms "recombination region" or "targeting sequence" refer to a segment (i.e., a portion) of a gene activation construct having a sequence that is substantially identical to or substantially complementary to a genomic gene sequence, e.g., including 5' flanking sequences of the genomic gene, and can facilitate homologous recombination between the genomic sequence and the targeting transgene construct.

[0115] As used herein, the term "replacement region" refers to a portion of a activation construct which becomes integrated into an endogenous chromosomal location following homologous recombination between a recombination region and a genomic sequence.

[0116] The heterologous regulatory sequences, e.g., which are provided in the replacement region, can include one or more of a variety elements, including: promoters (such as constitutive or inducible promoters), enhancers, negative regulatory elements, locus control regions, transcription factor binding sites, or combinations thereof.

[0117] Promoters/enhancers which may be used to control expression of the targeted gene in vivo include, but are not limited to, the cytomegalovirus (CMV) promoter/enhancer (Karasuyama et al., 1989, *J. Exp. Med.,* 169:13), the human β-actin promoter (Gunning et al. (1987) *PNAS* 84:4831-4835), the glucocorticoid-inducible promoter present in the mouse mammary tumor virus long terminal repeat (MMTV LTR) (Klessig et al. (1984) *Mol. Cell Biol.* 4:1354-1362), the long terminal repeat sequences of Moloney murine leukemia virus (MuLV LTR) (Weiss et al. (1985) *RNA Tumor Viruses*, Cold Spring Harbor Laboratory, Cold Spring Harbor, N.Y.), the SV40 early or late region promoter (Bernoist et al. (1981) *Nature* 290:304-310; Templeton et al. (1984) *Mol. Cell Biol.*, 4:817; and Sprague et al. (1983) *J. Virol,* 45:773), the promoter contained in the 3' long terminal repeat of Rous sarcoma virus (RSV) (Yamamoto et al., 1980, *Cell,* 22:787-797), the herpes simplex virus (HSV) thymidine kinase promoter/enhancer (Wagner et al. (1981) *PNAS* 82:3567-71), and the herpes simplex virus LAT promoter (Wolfe et al. (1992) *Nature Genetics,* 1:379-384).

[0118] In still other embodiments, the replacement region merely deletes a negative transcriptional control element of the native gene, e.g., to activate expression, or ablates a positive control element, e.g., to inhibit expression of the targeted gene.

[0119] B. Cell/Organism

[0120] The cell with the target gene may be derived from or contained in any organism (e.g., plant, animal, protozoan, virus, bacterium, or fungus). The dsRNA construct may be synthesized either in vivo or in vitro. Endogenous RNA polymerase of the cell may mediate transcription in vivo, or cloned RNA polymerase can be used for transcription in vivo or in vitro. For generating double stranded transcripts from a transgene in vivo, a regulatory region may be used to transcribe the RNA strand (or strands).

[0121] Furthermore, genetic manipulation becomes possible in organisms that are not classical genetic models. Breeding and screening programs may be accelerated by the ability to rapidly assay the consequences of a specific, targeted gene disruption. Gene disruptions may be used to discover the function of the target gene, to produce disease models in which the target gene are involved in causing or preventing a pathological condition, and to produce organisms with improved economic properties.

[0122] The cell with the target gene may be derived from or contained in any organism. The organism may a plant, animal, protozoan, bacterium, virus, or fungus. The plant may be a monocot, dicot or gymnosperm; the animal may be a vertebrate or invertebrate. Preferred microbes are those used in agriculture or by industry, and those that are pathogenic for plants or animals. Fungi include organisms in both the mold and yeast morphologies.

[0123] Plants include arabidopsis; field crops (e.g., alfalfa, barley, bean, corn, cotton, flax, pea, rape, rice, rye, safflower, sorghum, soybean, sunflower, tobacco, and wheat); vegetable crops (e.g., asparagus, beet, broccoli, cabbage, carrot, cauliflower, celery, cucumber, eggplant, lettuce, onion, pepper, potato, pumpkin, radish, spinach, squash, taro, tomato, and zucchini); fruit and nut crops (e.g., almond, apple, apricot, banana, blackberry, blueberry, cacao, cherry, coconut, cranberry, date, faJoa, filbert, grape, grapefruit, guava, kiwi, lemon, lime, mango, melon, nectarine, orange, papaya, passion fruit, peach, peanut, pear, pineapple, pistachio, plum, raspberry, strawberry, tangerine, walnut, and watermelon); and ornamentals (e.g., alder, ash, aspen, azalea, birch, boxwood, camellia, carnation, chrysanthemum, elm, fir, ivy, jasmine, juniper, oak, palm, poplar, pine, redwood, rhododendron, rose, and rubber).

[0124] Examples of vertebrate animals include fish, mammal, cattle, goat, pig, sheep, rodent, hamster, mouse, rat, primate, and human.

[0125] Invertebrate animals include nematodes, other worms, drosophila, and other insects. Representative generae of nematodes include those that infect animals (e.g., Ancylostoma, Ascaridia, Ascaris, Bunostomum, Caenorhabditis, Capillaria, Chabertia, Cooperia, Dictyocaulus, Haemonnchus, Heterakis, Nematodirus, Oesophagostomum, Ostertagia, Oxyuris, Parascaris, Strongylus, Toxascaris, Trichuris, Trichostrongylus, Tflichonema, Toxocara, Uncinaria) and those that infect plants (e.g., *B ursaphalenchus*, Criconerriella, Diiylenchus, Ditylenchus, Globodera, Helicotylenchus, Heterodera, Longidorus, Melodoigyne, Nacobbus, Paratylenchus, Pratylenchus, Radopholus, Rotelynchus, Tylenchus, and Xiphinerna). Representative orders of insects include Coleoptera, Diptera, Lepidoptera, and Homoptera.

[0126] The cell having the target gene may be from the germ line or somatic, totipotent or pluripotent, dividing or non-dividing, parenchyma or epithelium, immortalized or transformed, or the like. The cell may be a stem cell or a differentiated cell. Cell types that are differentiated include adipocytes, fibroblasts, myocytes, cardiomyocytes, endothelium, neurons, glia, blood cells, megakaryocytes, lymphocytes, macrophages, neutrophils, eosinophils, basophils, mast cells, leukocytes, granulocytes, keratinocytes, chondrocytes, osteoblasts, osteoclasts, hepatocytes, and cells of the endocrine or exocrine glands.

[0127] C. Targeted Genes

[0128] The target gene may be a gene derived from the cell, an endogenous gene, a transgene, or a gene of a pathogen which is present in the cell after infection thereof. Depending on the particular target gene and the dose of double stranded RNA material delivered, the procedure may provide partial or complete loss of function for the target gene. Lower doses of injected material and longer times after administration of dsRNA may result in inhibition in a smaller fraction of cells. Quantitation of gene expression in a cell may show similar amounts of inhibition at the level of accumulation of target mRNA or translation of target protein.

[0129] "Inhibition of gene expression" refers to the absence (or observable decrease) in the level of protein and/or mRNA product from a target gene. "Specificity" refers to the ability to inhibit the target gene without manifest effects on other genes of the cell. The consequences of inhibition can be confirmed by examination of the outward properties of the cell or organism (as presented below in the examples) or by biochemical techniques such as RNA solution hybridization, nuclease protection, Northern hybridization, reverse transcription, gene expression monitoring with a microarray, antibody binding, enzyme linked immunosorbent assay (ELISA), Western blotting, radioimmunoassay (RIA), other immunoassays, and fluorescence activated cell analysis (FACS). For RNA-mediated inhibition in a cell line or whole organism, gene expression is conveniently assayed by use of a reporter or drug resistance gene whose protein product is easily assayed. Such reporter genes include acetohydroxyacid synthase (AHAS), alkaline phosphatase (AP), beta galactosidase (LacZ), beta glucoronidase (GUS), chloramphenicol acetyltransferase (CAT), green fluorescent protein (GFP), horseradish peroxidase (HRP), luciferase (Luc), nopaline synthase (NOS), octopine synthase (OCS), and derivatives thereof multiple selectable markers are available that confer resistance to ampicillin, bleomycin, chloramphenicol, gentamycin, hygromycin, kanamycin, lincomycin, methotrexate, phosphinothricin, puromycin, and tetracyclin.

[0130] Depending on the assay, quantitation of the amount of gene expression allows one to determine a degree of inhibition which is greater than 10%, 33%, 50%, 90%, 95% or 99% as compared to a cell not treated according to the present invention. Lower doses of injected material and longer times after administration of dsRNA may result in inhibition in a smaller fraction of cells (e.g., at least 10%, 20%, 50%, 75%,90%, or 95% of targeted cells). Quantitation of gene expression in a cell may show similar amounts of inhibition at the level of accumulation of target mRNA or translation of target protein. As an example, the efficiency of inhibition may be determined by assessing the amount of gene product in the cell: mRNA may be detected with a hybridization probe having a nucleotide sequence outside the region used for the inhibitory double-stranded RNA, or translated polypeptide may be detected with an antibody raised against the polypeptide sequence of that region.

[0131] As disclosed herein, the present invention may is not limited to any type of target gene or nucleotide sequence. But the following classes of possible target genes are listed for illustrative purposes: developmental genes (e.g., adhesion molecules, cyclin kinase inhibitors, Writ family members, Pax family members, Winged helix family members, Hox family members, cytokines/lymphokines and their receptors, growth/differentiation factors and their receptors, neurotransmitters and their receptors); oncogenes (e.g., ABLI, BCLI, BCL2, BCL6, CBFA2, CBL, CSFIR, ERBA, ERBB, EBRB2, ETSI, ETSI, ETV6, FGR, FOS, FYN, HCR, HRAS, JUN, KRAS, LCK, LYN, MDM2, MLL, MYB, MYC, MYCLI, MYCN, NRAS, PIM 1, PML, RET, SRC, TALI, TCL3, and YES); tumor suppressor genes (e.g., APC, BRCA 1, BRCA2, MADH4, MCC, NF 1, NF2, RB 1, TP53, and WTI); and enzymes (e.g., ACC synthases and oxidases, ACP desaturases and hydroxylases, ADP-glucose pyrophorylases, ATPases, alcohol dehydrogenases, amylases, amyloglucosidases, catalases, cellulases, chalcone synthases, chitinases, cyclooxygenases, decarboxylases, dextrinases, DNA and RNA polymerases, galactosidases, glucanases, glucose oxidases, granule-bound starch synthases, GTPases, helicases, hemicellulases, integrases, inulinases, invertases, isomerases, kinases, lactases, lipases, lipoxygenases, lysozymes, nopaline synthases, octopine synthases, pectinesterases, peroxidases, phosphatases, phospholipases, phosphorylases, phytases, plant growth regulator synthases, polygalacturonases, proteinases and peptidases, pullanases, recombinases, reverse transcriptases, RUBISCOs, topoisomerases, and xylanases).

[0132] D. dsRNA constructs

[0133] The dsRNA construct may comprise one or more strands of polymerized ribonucleotide. It may include modifications to either the phosphate-sugar backbone or the nucleoside. For example, the phosphodiester linkages of natural RNA may be modified to include at least one of a nitrogen or sulfur heteroatom. Modifications in RNA structure may be tailored to allow specific genetic inhibition

while avoiding a general panic response in some organisms which is generated by dsRNA. Likewise, hases may be modified to hlock the activity of adenosine deaminase. The dsRNA construct may he produced enzymatically or hy partial/total organic synthesis, any modified ribonucleotide can he introduced hy in vitro enzymatic or organic synthesis.

[0134] The dsRNA construct may he directly introduced into the cell (i.e., intracellularly); or introduced extracellularly into a cavity, interstitial space, into the circulation of an organism, introduced orally, or may be introduced hy bathing an organism in a solution containing RNA. Methods for oral introduction include direct mixing of RNA with food of the organism, as well as engineered approaches in which a species that is used as food is engineered to express an RNA, then fed to the organism to he affected. Physical methods of introducing nucleic, acids include injection directly into the cell or extracellular injection into the organism of an RNA solution.

[0135] The double-stranded structure may he formed hy a single self-complementary RNA strand or two complementary RNA strands. RNA duplex formation may he initiated either inside or outside the cell. The RNA may be introduced in an amount which allows delivery of at least one copy per cell. Higher doses (e.g., at least 5, 10, 100, 500 or 1000 copies per cell) of double-stranded material may yield more effective inhihition; lower doses may also he useful for specific applications. Inhihition is sequence-specific in that nucleotide sequences corresponding to the duplex region of the RNA are targeted for genetic inhihition.

[0136] dsRNA constructs containing a nucleotide sequences identical to a portion of the target gene are preferred for inhibition. RNA sequences with insertions, deletions, and single point mutations relative to the target sequence have also been found to be effective for inhibition. Thus, sequence identity may optimized by sequence comparison and alignment algorithms known in the art (see Grihskov and Devereux, Sequence Analysis Primer, Stockton Press, 1991, and references cited therein) and calculating the percent difference between the nucleotide sequences by, for example, the Smith-Waterman algorithm as implemented in the BESTFIT software program using default parameters (e.g., University of Wisconsin Genetic Computing Group). Greater than 90% sequence identity, or even 100% sequence identity, between the inhihitory RNA and the portion of the target gene is preferred. Alternatively, the duplex region of the RNA may he defined functionally as a nucleotide sequence that is capahle of hyhridizing with a portion of the target gene transcript (e.g., 400 mM NaCl, 40 mM PIPES pH 6.4, 1 mM EDTA, 50° C. or 70° C. hyhridization for 12-16 hours; followed by washing). In certain preferred embodiments, the length of the dsRNA is at least 20, 21 or 22 nucleotides in length, e.g., corresponding in size to RNA products produced by Dicer-dependent cleavage. In certain emhodiments, the dsRNA construct is at least 25, 50, 100, 200, 300 or 400 bases. In certain emhodiments, the dsRNA construct is 400-800 hases in length.

[0137] 100% sequence identity between the RNA and the target gene is not required to practice the present invention. Thus the invention has the advantage of heing able to tolerate sequence variations that might he expected due to genetic mutation, strain polymorphism, or evolutionary divergence.

[0138] The dsRNA construct may he synthesized either in vivo or in vitro. Endogenous RNA polymerase of the cell may mediate transcription in vivo, or cloned RNA polymerase can be used for transcription in vivo or in vitro. For transcription from a transgene in vivo or an expression construct, a regulatory region (e.g., promoter, enhancer, silencer, splice donor and acceptor, polyadenylation) may be used to transcrihe the dsRNA strand (or strands). Inhihition may he targeted by specific transcription in an organ, tissue, or cell type; stimulation of an environmental condition (e.g., infection, stress, temperature, chemical inducers); and/or engineering transcription at a developmental stage or age. The RNA strands may or may not be polyadenylated; the RNA strands may or may not be capahle of being translated into a polypeptide by a cell's translational apparatus. The dsRNA construct may be chemically or enzymatically synthesized hy manual or automated reactions. The dsRNA construct may he synthesized hy a cellular RNA polymerase or a hacteriophage RNA polymerase (e.g., T3, T7, SP6). The use and production of an expression construct are known in the art 32,33,34 (see also WO 97/32016; U.S. Pat. Nos. 5,593,874, 5,698,425, 5,712,135, 5,789,214, and 5,804,693; and the references cited therein). If synthesized chemically or hy in vitro enzymatic synthesis, the RNA may be purified prior to introduction into the cell. For example, RNA can he punified from a mixture hy extraction with a solvent or resin, precipitation, electrophoresis, chromatography or a comhination thereof. Alternatively, the dsRNA construct may he used with no or a minimum of purification to avoid losses due to sample processing. The dsRNA construct may he dried for storage or dissolved in an aqueous solution. The solution may contain huffers or salts to promote annealing, and/or stahilization of the duplex strands.

[0139] Physical methods of introducing nucleic acids include injection of a solution containing the dsRNA construct, hombardment by particles covered hy the dsRNA construct, soaking the cell or organism in a solution of the RNA, or electroporation of cell membranes in the presence of the dsRNA construct. A viral construct packaged into a viral particle would accomplish hoth efficient introduction of an expression construct into the cell and transcription of dsRNA construct encoded by the expression construct. Other methods known in the art for introducing nucleic acids to cells may be used, such as lipid-mediated carrier transport, chemicalmediated transport, such as calcium phosphate, and the like. Thus the dsRNA construct may be introduced along with components that perform one or more of the following activities: enhance RNA uptake hy the cell, promote annealing of the duplex strands, stahilize the annealed strands, or other-wise increase inhihition of the target gene.

[0140] E. Illustrative Uses

[0141] One utility of the present invention is as a method of identifying gene function in an organism, especially higher eukaryotes comprising the use of double-stranded RNA to inhibit the activity of a target gene of previously unknown function. Instead of the time consuming and lahorious isolation of mutants by traditional genetic screening, functional genomics would envision determining the function of uncharacterized genes hy employing the invention to reduce the amount and/or alter the timing of target gene activity. The invention could be used in determining potential targets for pharmaceutics, understanding normal

and pathological events associated with development, determining signaling pathways responsible for postnatal development/aging, and the like. The increasing speed of acquiring nucleotide sequence information from genomic and expressed gene sources, including total sequences for mammalian genomes, can be coupled with the invention to determine gene function in a cell or in a whole organism. The preference of different organisms to use particular codons, searching sequence databases for related gene products, correlating the linkage map of genetic traits with the physical map from which the nucleotide sequences are derived, and artificial intelligence methods may be used to define putative open reading frames from the nucleotide sequences acquired in such sequencing projects.

[0142] A simple assay would be to inhibit gene expression according to the partial sequence available from an expressed sequence tag (EST). Functional alterations in growth, development, metabolism, disease resistance, or other biological processes would be indicative of the normal role of the EST's gene product.

[0143] The ease with which the dsRNA construct can he introduced into an intact cell/organism containing the target gene allows the present invention to be used in high throughput screening (HTS). For example, duplex RNA can be produced by an amplification reaction using primers flanking the inserts of any gene library derived from the target cell or organism. Inserts may be derived from genomic DNA or mRNA (e.g., cDNA and cRNA). Individual clones from the library can be replicated and then isolated in separate reactions, but preferably the library is maintained in individual reaction vessels (e.g., a 96 well microtiter plate) to minimize the number of steps required to practice the invention and to allow automation of the process.

[0144] In an exemplary embodiment, the subject invention provides an arrayed library of RNAi constructs. The array may in the form of solutions, such as multi-well plates, or may be "printed" on solid substrates upon which cells can be grown. To illustrate, solutions containing duplex RNAs that are capable of inhibiting the different expressed genes can be placed into individual wells positioned on a microtiter plate as an ordered array, and intact cells/organisms in each well can be assayed for any changes or modifications in behavior or development due to inhibition of target gene activity.

[0145] In one embodiment, the subject method uses an arrayed library of RNAi constructs to screen for combinations of RNAi that is lethal to host cells. Synthetic lethality is a bedrock principle of experimental genetics. A synthetic lethality describes the properties of two mutations which, individually, are tolerated hy the organism hut which, in combination, are lethal. The subject arrays can he used to identify loss-of-function mutations that are lethal in combination with alterations in other genes, such as activated oncogenes or loss-of-function mutations to tumor suppressors. To achieve this, one can create "phenotype arrays" using cultured cells. Expression of each of a set of genes, such as the host cell's genome, can be individually systematically disrupted using RNA interference. Combination with alterations in oncogene and tumor suppressor pathways can he used to identify synthetic lethal interactions that may identify novel therapeutic targets.

[0146] In certain embodiments, the RNAi constructs can be fed directly to, injected into, the cell/organism containing the target gene. Alternatively, the duplex RNA can he produced hy in vivo or in vitro transcription from an expression construct used to produce the library. The construct can be replicated as individual clones of the library and transcribed to produce the RNA; each clone can then be fed to, or injected into, the cell/organism containing the target gene. The function of the target gene can be assayed from the effects it has on the cell/organism when gene activity is inhibited. This screening could be amenable to small subjects that can be processed in large number, for example, tissue culture cells derived from mammals, especially primates, and most preferably humans.

[0147] If a characteristic of an organism is determined to be genetically linked to a polymorphism through RFLP or QTL analysis, the present invention can be used to gain insight regarding whether that genetic polymorphism might be directly responsible for the characteristic. For example, a fragment defining the genetic polymorphism or sequences in the vicinity of such a genetic polymorphism can be amplified to produce an RNA, the duplex RNA can be introduced to the organism or cell, and whether an alteration in the charactenstic is correlated with inhibition can he determined. Of course, there may be trivial explanations for negative results with this type of assay, for example: inhibition of the target gene causes lethality, inhibition of the target gene may not result in any observable alteration, the fragment contains nucleotide sequences that are not capable of inhibiting the target gene, or the target gene's activity is redundant.

[0148] The present invention may be useful in allowing the inhibition of essential genes. Such genes may be required for cell or organism viability at only particular stages of development or cellular compartments. The functional equivalent of conditional mutations may be produced by inhibiting activity of the target gene when or where it is not required for viability. The invention allows addition of RNA at specific times of development and locations in the organism without introducing permanent mutations into the target genome.

[0149] If alternative splicing produced a family of transcripts that were distinguished hy usage of characteristic exons, the present invention can target inhibition through the appropriate exons to specifically inhibit or to distinguish among the functions of family members. For example, a hormone that contained an alternatively spliced transmembrane domain may be expressed in both membrane hound and secreted forms. Instead of isolating a nonsense mutation that terminates translation before the transmemhrane domain, the functional consequences of having only secreted hormone can he determined according to the invention hy targeting the exon containing the transmembrane domain and thereby inhibiting expression of membrane-hound hormone.

[0150] The present invention may be used alone or as a component of a kit having at least one of the reagents necessary to carry out the in vitro or in vivo introduction of RNA to test samples or subjects. Preferred components are the dsRNA and a vehicle that promotes introduction of the dsRNA. Such a kit may also include instructions to allow a user of the kit to practice the invention.

[0151] Alternatively, an organism may be engineered to produce dsRNA which produces commercially or medically

beneficial results, for example, resistance to a pathogen or its pathogenic effects, improved growth, or novel developmental patterns.

IV. Exemplification

[0152] The invention, now being generally described, will be more readily understood by reference to the following examples, which are included merely for purposes of illustration of certain aspects and embodiments of the present invention and are not intended to limit the invention.

Example 1

An RNA-directed Nuclease Mediates RNAi Gene Silencing

[0153] In a diverse group of organisms that includes *Caenorhabditis elegans*, Drosophila, planaria, hydra, trypanosomes, fungi and plants, the introduction of double-stranded RNAs inhibits gene expression in a sequence-specific manner[1-7]. These responses, called RNA interference or post-transcriptional gene silencing, may provide anti-viral defence, modulate transposition or regulate gene expression[2, 6, 8-10]. We have taken a biochemical approach towards elucidating the mechanisms underlying this genetic phenomenon. Here we show that 'loss-of-function' phenotypes can be created in cultured Drosophila cells by transfection with specific double-stranded RNAs. This coincides with a marked reduction in the level of cognate cellular messenger RNAs. Extracts of transfected cells contain a nuclease activity that specifically degrades exogenous transcripts homologous to transfected double-stranded RNA. This enzyme contains an essential RNA component. After partial purification, the sequence-specific nuclease co-fractionates with a discrete, ~25-nucleotide RNA species which may confer specificity to the enzyme through homology to the substrate mRNAs.

[0154] Although double-stranded RNAs (dsRNAs) can provoke gene silencing in numerous biological contexts including Drosophila[11, 12], the mechanisms underlying this phenomenon have remained mostly unknown. We therefore wanted to establish a biochemically tractable model in which such mechanisms could be investigated.

[0155] Transient transfection of cultured, Drosophila S2 cells with a lacZ expression vector resulted in β-galactosidase activity that was easily detectable by an in situ assay (**FIG. 1***a*). This activity was greatly reduced by co-transfection with a dsRNA corresponding to the first 300 nucleotides of the lacZ sequence, whereas co-transfection with a control dsRNA (CD8) (**FIG. 1***a*) or with single-stranded RNAs of either sense or antisense orientation (data not shown) had little or no effect. This indicated that dsRNAs could interfere, in a sequence-specific fashion, with gene expression in cultured cells.

[0156] To determine whether RNA interference (RNAi) could be used to target endogenous genes, we transfected S2 cells with a dsRNA corresponding to the first 540 nucleotides of Drosophila cyclin E, a gene that is essential for progression into S phase of the cell cycle. During log-phase growth, untreated S2 cells reside primarily in G2/M (**FIG. 1***b*). Transfection with lacZ dsRNA had no effect on cell-cycle distribution, but transfection with the cyclin E dsRNA caused a G1-phase cell-cycle arrest (**FIG. 1***b*). The ability of cyclin E dsRNA to provoke this response was length-dependent. Double-stranded RNAs of 540 and 400 nucleotides were quite effective, whereas dsRNAs of 200 and 300 nucleotides were less potent. Double-stranded cyclin E RNAs of 50 or 100 nucleotides were inert in our assay, and transfection with a single-stranded, antisense cyclin E RNA had virtually no effect.

[0157] One hallmark of RNAi is a reduction in the level of mRNAs that are homologous to the dsRNA. Cells transfected with the cyclin E dsRNA (bulk population) showed diminished endogenous cyclin E mRNA as compared with control cells (**FIG. 1***c*). Similarly, transfection of cells with dsRNAs homologous to fizzy, a component of the anaphase-promoting complex (APC) or cyclin A, a cyclin that acts in S, G2 and M, also caused reduction of their cognate mRNAs (**FIG. 1***c*). The modest reduction in fizzy mRNA levels in cells transfected with cyclin A dsRNA probably resulted from arrest at a point in the division cycle at which fizzy transcription is low[14, 15]. These results indicate that RNAi may be a generally applicable method for probing gene function in cultured Drosophila cells.

[0158] The decrease in mRNA levels observed upon transfection of specific dsRNAs into Drosophila cells could be explained by effects at transcriptional or post-transcriptional levels. Data from other systems have indicated that some elements of the dsRNA response may affect mRNA directly (reviewed in refs **1** and **6**). We therefore sought to develop a cell-free assay that reflected, at least in part, RNAi.

[0159] S2 cells were transfected with dsRNAs corresponding to either cyclin E or lacZ. Cellular extracts were incubated with synthetic mRNAs of lacZ or cyclin E. Extracts prepared from cells transfected with the 540-nucleotide cyclin E dsRNA efficiently degraded the cyclin E transcript; however, the lacZ transcript was stable in these lysates (**FIG. 2***a*). Conversely, lysates from cells transfected with the lacZ dsRNA degraded the lacZ transcript but left the cyclin E mRNA intact. These results indicate that RNAi ablates target mRNAs through the generation of a sequence-specific nuclease activity. We have termed this enzyme RISC (RNA-induced silencing complex). Although we occasionally observed possible intermediates in the degradation process (see **FIG. 2**), the absence of stable cleavage end-products indicates an exonuclease (perhaps coupled to an endonuclease). However, it is possible that the RNAi nuclease makes an initial endonucleolytic cut and that non-specific exonucleases in the extract complete the degradation process[16]. In addition, our ability to create an extract that targets lacZ in vitro indicates that the presence of an endogenous gene is not required for the RNAi response.

[0160] To examine the substrate requirements for the dsRNA-induced, sequence-specific nuclease activity, we incubated a variety of cyclin-E-derived transcripts with an extract derived from cells that had been transfected with the 540-nucleotide cyclin E dsRNA (**FIG. 2***b*, *c*). Just as a length requirement was observed for the transfected dsRNA, the RNAi nuclease activity showed a dependence on the size of the RNA substrate. Both a 600-nucleotide transcript that extends slightly beyond the targeted region (**FIG. 2***b*) and an ~1-kilobase (kb) transcript that contains the entire coding sequence (data not shown) were completely destroyed by the extract. Surprisingly, shorter substrates were not degraded as

efficiently. Reduced activity was observed against either a 300- or a 220-nucleotide transcript, and a 100-nucleotide transcript was resistant to nuclease in our assay. This was not due solely to position effects because ~100-nucleotide transcripts derived from other portions of the transfected dsRNA behaved similarly (data not shown). As expected, the nuclease activity (or activities) present in the extract could also recognize the antisense strand of the cyclin E mRNA. Again, substrates that contained a substantial portion of the targeted region were degraded efficiently whereas those that contained a shorter stretch of homologous sequence (~130 nucleotides) were recognized inefficiently (**FIG. 2***c*, as600). For both the sense and antisense strands, transcripts that had no homology with the transfected dsRNA (**FIG. 2***b*, Eout; **FIG. 2***c*, as300) were not degraded. Although we cannot exclude the possibility that nuclease specificity could have migrated beyond the targeted region, the resistance of transcripts that do not contain homology to the dsRNA is consistent with data from *C. elegans*. Double-stranded RNAs homologous to an upstream cistron have little or no effect on a linked downstream cistron, despite the fact that unprocessed, polycistronic mRNAs can be readily detected[17]. Furthermore, the nuclease was inactive against a dsRNA identical to that used to provoke the RNAi response in vivo (**FIG. 2***b*). In the in vitro system, neither a 5' cap nor a poly(A) tail was required, as such transcripts were degraded as efficiently as uncapped and non-polyadenylated RNAs.

[0161] Gene silencing provoked by dsRNA is sequence specific. A plausible mechanism for determining specificity would be incorporation of nucleic-acid guide sequences into the complexes that accomplish silencing[19]. In accord with this idea, pre-treatment of extracts with a Ca^{2+}-dependent nuclease (micrococcal nuclease) abolished the ability of these extracts to degrade cognate mRNAs (**FIG. 3**). Activity could not be rescued by addition of non-specific RNAs such as yeast transfer RNA. Although micrococcal nuclease can degrade both DNA and RNA, treatment of the extract with DNAse I had no effect (**FIG. 3**). Sequence-specific nuclease activity, however, did require protein (data not shown). Together, our results support the possibility that the RNAi nuclease is a ribonucleoprotein, requiring both RNA and protein components. Biochemical fractionation (see below) is consistent with these components being associated in extract rather than being assembled on the target mRNA after its addition.

[0162] In plants, the phenomenon of co-suppression has been associated with the existence of small (~25-nucleotide) RNAs that correspond to the gene that is being silenced[19]. To address the possibility that a similar RNA might exist in Drosophila and guide the sequence-specific nuclease in the choice of substrate, we partially purified our activity through several fractionation steps. Crude extracts contained both sequence-specific nuclease activity and abundant, heterogeneous RNAs homologous to the transfected dsRNA (**FIGS. 2 and 4***a*). The RNAi nuclease fractionated with ribosomes in a high-speed centrifugation step. Activity could be extracted by treatment with high salt, and ribosomes could be removed by an additional centrifugation step. Chromatography of soluble nuclease over an anion-exchange column resulted in a discrete peak of activity (**FIG. 4***b*, cyclin E). This retained specificity as it was inactive against a heterologous mRNA (**FIG. 4***b*, lacZ). Active fractions also contained an RNA species of 25 nucleotides that is homologous to the cyclin E target (**FIG. 4***b*, northern). The band observed on northern blots may represent a family of discrete RNAs because it could be detected with probes specific for both the sense and antisense cyclin E sequences and with probes derived from distinct segments of the dsRNA (data not shown). At present, we cannot determine whether the 25-nucleotide RNA is present in the nuclease complex in a double-stranded or single-stranded form.

[0163] RNA interference allows an adaptive defence against both exogenous and endogenous dsRNAs, providing something akin to a dsRNA immune response. Our data, and that of others[19], is consistent with a model in which dsRNAs present in a cell are converted, either through processing or replication, into small specificity determinants of discrete size in a manner analogous to antigen processing. Our results suggest that the post-transcriptional component of dsRNA-dependent gene silencing is accomplished by a sequence-specific nuclease that incorporates these small RNAs as guides that target specific messages based upon sequence recognition. The identical size of putative specificity determinants in plants[19] and animals predicts a conservation of both the mechanisms and the components of dsRNA-induced, post-transcriptional gene silencing in diverse organisms. In plants, dsRNAs provoke not only post-transcriptional gene silencing but also chromatin remodelling and transcriptional repression[20, 21]. It is now critical to determine whether conservation of gene-silencing mechanisms also exists at the transcriptional level and whether chromatin remodelling can be directed in a sequence-specific fashion by these same dsRNA-derived guide sequences.

[0164] Methods

[0165] Cell Culture and RNA Methods

[0166] S2 (ref. 22) cells were cultured at 27° C. in 90% Schneider's insect media (Sigma), 10% heat inactivated fetal bovine serum (FBS). Cells were transfected with dsRNA and plasmid DNA by calcium phosphate co-precipitation[23]. Identical results were observed when cells were transfected using lipid reagents (for example, Superfect, Qiagen). For FACS analysis, cells were additionally transfected with a vector that directs expression of a green fluorescent protein (GFP)-US9 fusion protein[13]. These cells were fixed in 90% ice-cold ethanol and stained with propidium iodide at 25 $\mu g\ ml^{-1}$. FACS was performed on an Elite flow cytometer (Coulter). For northern blotting, equal loading was ensured by over-probing blots with a control complementary DNA (RP49). For the production of dsRNA, transcription templates were generated by polymerase chain reaction such that they contained T7 promoter sequences on each end of the template. RNA was prepared using the RiboMax kit (Promega). Confirmation that RNAs were double stranded came from their complete sensitivity to RNAse III (a gift from A. Nicholson). Target mRNA transcripts were synthesized using the Riboprobe kit (Promega) and were gel purified before use.

[0167] Extract Preparation

[0168] Log-phase S2 cells were plated on 15-cm tissue culture dishes and transfected with 30 μg dsRNA and 30 μg carrier plasmid DNA. Seventy-two hours after transfection, cells were harvested in PBS containing 5 mM EGTA washed twice in PBS and once in hypotonic buffer (10 mM HEPES pH 7.3, 6 mM β-mercaptoethanol). Cells were suspended in

0.7 packed-cell volumes of hypotonic buffer containing Complete protease inhibitors (Boehringer) and 0.5 units ml^{-1} of RNasin (Promega). Cells were disrupted in a dounce homogenizer with a type B pestle, and lysates were centrifuged at 30,000 g for 20 min. Supernatants were used in an in vitro assay containing 20 mM HEPES pH 7.3, 110 mM KOAc, 1 mM $Mg(OAc)_2$, 3 mM EGTA, 2 mM $CaCl_2$, 1 mM DTT. Typically, 5 μl extract was used in a 10 μl assay that contained also 10,000 c.p.m. synthetic mRNA substrate.

[0169] Extract Fractionation

[0170] Extracts were centrifuged at 200,000 g for 3 h and the resulting pellet (containing ribosomes) was extracted in hypotonic buffer containing also 1 mM $MgCl_2$ and 300 mM KOAc. The extracted material was spun at 100,000 g for 1 h and the resulting supernatant was fractionated on Source 15Q column (Pharmacia) using a KCl gradient in buffer A (20 mM HEPES pH 7.0, 1 mM dithiothreitol, 1 mM $MgCl_2$). Fractions were assayed for nuclease activity as described above. For northern blotting, fractions were proteinase K/SDS treated, phenol extracted, and resolved on 15% acrylamide 8M urea gels. RNA was electroblotted onto Hybond N+ and probed with strand-specific riboprobes derived from cyclin E mRNA. Hybridization was carried out in 500 mM $NaPO_4$ pH 7.0, 15% formamide, 7% SDS, 1% BSA. Blots were washed in 1 SSC at 37-45° C.

References Cited in Example 1

[0171] 1. Sharp, P. A. RNAi and double-strand RNA. *Genes Dev.* 13, 139-141 (1999).

[0172] 2. Sanchez-Alvarado, A. & Newmark, P. A. Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc. Natl Acad. Sci. USA* 96, 5049-5054 (1999).

[0173] 3. Lohmann, J. U., Endl, I. & Bosch, T. C. Silencing of developmental genes in Hydra. *Dev. Biol.* 214, 211-214 (1999).

[0174] 4. Cogoni, C. & Macino, G. Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase. *Nature* 399, 166-169 (1999).

[0175] 5. Waterhouse, P. M., Graham, M. W. & Wang, M. B. Virus resistance and gene silencing in plants can be induced by simultaneous expression of sense and antisense RNA. *Proc. Natl Acad. Sci. USA* 95, 13959-13964 (1998).

[0176] 6. Montgomery, M. K. & Fire, A. Double-stranded RNA as a mediator in sequence-specific genetic silencing and co-suppression. *Trends Genet.* 14, 225-228 (1998).

[0177] 7. Ngo, H., Tschudi, C., Gull, K. & Ullu, E. Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. *Proc. Natl Acad. Sci. USA* 95, 14687-14692 (1998).

[0178] 8. Tabara, H. et al. The rde-1 gene, RNA interference, and transposon silencing in *C. elegans. Cell* 99, 123-132 (1999).

[0179] 9. Ketting, R. F., Haverkamp, T. H. A., van Luenen, H. G. A. M. & Plasterk, R. H. A. mut-7 of *C. elegans,* required for transposon silencing and RNA interference, is a homolog of Werner Syndrome helicase and RnaseD. *Cell* 99, 133-141 (1999).

[0180] 10. Ratcliff, F., Harrison, B. D. & Baulcombe, D. C. A similarity between viral defense and gene silencing in plants. *Science* 276, 1558-1560 (1997).

[0181] 11. Kennerdell, J. R. & Carthew, R. W. Use of dsRNA-mediated genetic interference to demonstrate that frizzled and frizzled 2 act in the wingless pathway. *Cell* 95, 1017-1026 (1998).

[0182] 12. Misquitta, L. & Paterson, B. M. Targeted disruption of gene function in Drosophila by RNA interference: a role for nautilus in embryonic somatic muscle formation. *Proc. Natl Acad. Sci. USA* 96, 1451-1456 (1999).

[0183] 13. Kalejta, R. F., Brideau, A. D., Banfield, B. W. & Beavis, A. J. An integral membrane green fluorescent protein marker, Us9-GFP, is quantitatively retained in cells during propidium iodine-based cell cycle analysis by flow cytometry. *Exp. Cell. Res.* 248, 322-328 (1999).

[0184] 14. Wolf, D. A. & Jackson, P. K. Cell cycle: oiling the gears of anaphase. *Curr. Biol.* 8, R637-R639 (1998).

[0185] 15. Kramer, E. R., Gieffers, C., Holz, G., Hengstschlager, M. & Peters, J. M. Activation of the human anaphase-promoting complex by proteins of the CDC20/fizzy family. *Curr. Biol.* 8, 1207-1210 (1998).

[0186] 16. Shuttleworth, J. & Colman, A. Antisense oligonucleotide-directed cleavage of mRNA in Xenopus oocytes and eggs. *EMBO J.* 7, 427-434 (1988).

[0187] 17. Tabara, H., Grishok, A. & Mello, C. C. RNAi in *C. elegans:* soaking in the genome sequence. *Science* 282, 430-432 (1998).

[0188] 18. Bosher, J. M., Dufourcq, P., Sookhareea, S. & Labouesse, M. RNA interference can target pre-mRNA. Consequences for gene expression in a *Caenorhabditis elegans* operon. *Genetics* 153, 1245-1256 (1999).

[0189] 19. Hamilton, J. A. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science* 286, 950-952 (1999).

[0190] 20. Jones, L. A., Thomas, C. L. & Maule, A. J. De novo methylation and co-suppression induced by a cytoplasmically replicating plant RNA virus. *EMBO J.* 17, 6385-6393 (1998).

[0191] 21. Jones, L. A. et al. RNA-DNA interactions and DNA methylation in post-transcriptional gene silencing. *Plant Cell* 11, 2291-2301 (1999).

[0192] 22. Schneider, I. Cell lines derived from late embryonic stages of *Drosophila melanogaster. J. Embryol. Exp. Morpho.* 27, 353-365 (1972).

[0193] 23. Di Nocera, P. P. & Dawid, I. B. Transient expression of genes introduced into cultured cells of Drosophila. *Proc. Natl Acad. Sci. USA* 80, 7095-7098 (1983).

Example 2

Role for a Bidentate Ribonuclease in the Initiation Step of RNA Interference

[0194] Genetic approaches in worms, fungi and plants have identified a group of proteins that are essential for double-stranded RNA-induced gene silencing. Among these are ARGONAUTE family members (e.g. RDE1, QDE2)[9,10,30], recQ-family helicases (MUT-7, QDE3)[11,12], and RNA-dependent RNA polymerases (e.g. EGO-1, QDE1, SGS2/SDE1)[13-16]. While potential roles have been proposed, none of these genes has been assigned a definitive function in the silencing process. Biochemical studies have suggested that PTGS is accomplished by a multicomponent nuclease that targets mRNAs for degradation[6,8,17]. We have shown that the specificity of this complex may derive from the incorporation of a small guide sequence that is homologous to the mRNA substrate[6]. Originally identified in plants that were actively silencing transgenes[7], these ~22 nt. RNAs have been produced during RNAi in vitro using an extract prepared from Drosophila embryos[8]. Putative guide RNAs can also be produced in extracts from Drosophila S2 cells (**FIG. 5***a*). With the goal of understanding the mechanism of post-transcriptional gene silencing, we have undertaken both biochemical fractionation and candidate gene approaches to identify the enzymes that execute each step of RNAi.

[0195] Our previous studies resulted in the partial purification of a nuclease, RISC, that is an effector of RNA interference. See Example 1. This enzyme was isolated from Drosophila S2 cells in which RNAi had been initiated in vivo by transfection with dsRNA. We first sought to determine whether the RISC enzyme and the enzyme that initiates RNAi via processing of dsRNA into 22 mers are distinct activities. RISC activity could be largely cleared from extracts by high-speed centrifugation (100,000×g for 60 min.) while the activity that produces 22 mers remained in the supernatant (**FIG. 5***b,c*). This simple fractionation indicated that RISC and the 22 mer-generating activity are separable and thus distinct enzymes. However, it seems likely that they might interact at some point during the silencing process.

[0196] RNAse III family members are among the few nucleases that show specificity for double-stranded RNA[18]. Analysis of the Drosophila and *C. elegans* genomes reveals several types of RNAse III enzymes. First is the canonical RNAse III which contains a single RNAse III signature motif and a double-stranded RNA binding domain (dsRBD; e.g. RNC_CAEEL). Second is a class represented by Drosha[19], a Drosophila enzyme that contains two RNAse III motifs and a dsRBD (CeDrosha in *C. elegans*). A third class contains two RNAse III signatures and an amino terminal helicase domain (e.g. Drosophila CG4792, CG6493, *C. elegans* K12H4.8), and these had previously been proposed by Bass as candidate RNAi nucleases[20]. Representatives of all three classes were tested for the ability to produce discrete, ~22 nt. RNAs from dsRNA substrates.

[0197] Partial digestion of a 500 nt. cyclin E dsRNA with purified, bacterial RNAse III produced a smear of products while nearly complete digestion produced a heterogeneous group of ~11-17 nucleotide RNAs (not shown). In order to test the dual-RNAse III enzymes, we prepared T7 epitope-tagged versions of Drosha and CG4792. These were expressed in transfected S2 cells and isolated by immunoprecipitation using antibody-agarose conjugates. Treatment of the dsRNA with the CG4792 immunoprecipitate yielded ~22 nt. fragments similar to those produced in either S2 or embryo extracts (**FIG. 6***a*). Neither activity in extract nor activity in immunoprecipitates depended on the sequence of the RNA substrate since dsRNAs derived from several genes were processed equivalently (see Supplement 1). Negative results were obtained with Drosha and with immunoprecipitates of a DExH box helicase (Homeless[21]; see **FIGS. 6***a,b*). Western blotting confirmed that each of the tagged proteins was expressed and immunoprecipitated similarly (see Supplement 2). Thus, we conclude that CG4792 may carry out the initiation step of RNA interference by producing ~22 nt. guide sequences from dsRNAs. Because of its ability to digest dsRNA into uniformly sized, small RNAs, we have named this enzyme Dicer (Dcr). Dicer mRNA is expressed in embryos, in S2 cells, and in adult flies, consistent with the presence of functional RNAi machinery in all of these contexts (see Supplement 3).

[0198] The possibility that Dicer might be the nuclease responsible for the production of guide RNAs from dsRNAs prompted us to raise an antiserum directed against the carboxy-terminus of the Dicer protein (Dicer-1, CG4792). This antiserum could immunoprecipitate a nuclease activity from either Drosophila embryo extracts or from S2 cell lysates that produced ~22 nt. RNAs from dsRNA substrates (**FIG. 6***C*). The putative guide RNAs that are produced by the Dicer-1 enzyme precisely comigrate with 22 mers that are produced in extract and with 22 mers that are associated with the RISC enzyme (**FIGS. 6 D,F**). It had previously been shown that the enzyme that produced guide RNAs in Drosophila embryo extracts was ATP-dependent[8]. Depletion of this cofactor resulted in an ~6-fold lower rate of dsRNA cleavage and in the production of RNAs with a slightly lower mobility. Of interest was the fact that both Dicer-1 immunoprecipitates and extracts from S2 cells require ATP for the production of 22 mers (**FIG. 6D**). We do not observe the accumulation of lower mobility products in these cases, although we do routinely observe these in ATP-depleted embryo extracts. The requirement of this nuclease for ATP is a quite unusual property. We hypothesize that this requirement could indicate that the enzyme may act processively on the dsRNA, with the helicase domain harnessing the energy of ATP hydrolysis both for unwinding guide RNAs and for translocation along the substrate.

[0199] Efficient induction of RNA interference in *C. elegans* and in Drosophila has several requirements. For example, the initiating RNA must be double-stranded, and it must be several hundred nucleotides in length. To determine whether these requirements are dictated by Dicer, we characterized the ability of extracts and of immunoprecipitated enzyme to digest various RNA substrates. Dicer was inactive against single stranded RNAs regardless of length (see Supplement 4). The enzyme could digest both 200 and 500 nucleotide dsRNAs but was significantly less active with shorter substrates (see Supplement 4). Double-stranded RNAs as short as 35 nucleotides could be cut by the enzyme, albeit very inefficiently (data not shown). In contrast, *E. coli* RNAse III could digest to completion dsRNAs of 35 or 22 nucleotides (not shown). This suggests that the substrate preferences of the Dicer enzyme may contribute to but not wholly determine the size dependence of RNAi.

[0200] To determine whether the Dicer enzyme indeed played a role in RNAi in vivo, we sought to deplete Dicer activity from S2 cells and test the effect on dsRNA-induced gene silencing. Transfection of S2 cells with a mixture of dsRNAs homologous to the two Drosophila Dicer genes (CG4792 and CG6493) resulted in an ~6-7 fold reduction of Dicer activity either in whole cell lysates or in Dicer-1 immunoprecipitates (**FIGS. 7A,B**). Transfection with a control dsRNA (murine caspase 9) had no effect. Qualitatively similar results were seen if Dicer was examined by Northern blotting (not shown). Depletion of Dicer in this manner substantially compromised the ability of cells to silence subsequently an exogenous, GFP transgene by RNAi (**FIG. 7C**). These results indicate that Dicer is involved in RNAi in vivo. The lack of complete inhibition of silencing could result from an incomplete suppression of Dicer (which is itself required for RNAi) or could indicate that in vivo, guide RNAs can be produced by more than one mechanism (e.g. through the action of RNA-dependent RNA polymerases).

[0201] Our results indicate that the process of RNA interference can be divided into at least two distinct steps. According to this model, initiation of PTGS would occur upon processing of a double-stranded RNA by Dicer into ~22 nucleotide guide sequences, although we cannot formally exclude the possibility that another, Dicer-associated nuclease may participate in this process. These guide RNAs would be incorporated into a distinct nuclease complex (RISC) that targets single-stranded mRNAs for degradation. An implication of this model is that guide sequences are themselves derived directly from the dsRNA that triggers the response. In accord with this model, we have demonstrated that ^{32}P-labeled, exogenous dsRNAs that have been introduced into S2 cells by transfection are incorporated into the RISC enzyme as 22 mers (**FIG. 7E**). However, we cannot exclude the possibility that RNA-dependent RNA polymerases might amplify 22 mers once they have been generated or provide an alternative method for producing guide RNAs.

[0202] The structure of the Dicer enzyme provokes speculation on the mechanism by which the enzyme might produce discretely sized fragments irrespective of the sequence of the dsRNA (see Supplement 1, **FIG. 8a**). It has been established that bacterial RNAse III acts on its substrate as a dimer[18,22,23]. Similarly, a dimer of Dicer enzymes may be required for cleavage of dsRNAs into ~22 nt. pieces. According to one model, the cleavage interval would be determined by the physical arrangement of the two RNAse III domains within Dicer enzyme (**FIG. 8a**). A plausible alternative model would dictate that cleavage was directed at a single position by the two RIII domains in a single Dicer protein. The 22 nucleotide interval could be dictated by interaction of neighboring Dicer enzymes or by translocation along the mRNA substrate. The presence of an integral helicase domain suggests that the products of Dicer cleavage might be single-stranded 22 mers that are incorporated into the RISC enzyme as such.

[0203] A notable feature of the Dicer family is its evolutionary conservation. Homologs are found in *C. elegans* (K12H4.8), Arabidopsis (e.g., CARPEL FACTORY[24], T25K16.4, AC012328_1), mammals (Helicase-MOI[25]) and *S. pombe* (YC9A_SCHPO) (**FIG. 8b**, see Supplements 6,7 for sequence comparisons). In fact, the human Dicer family member is capable of generating ~22 nt. RNAs from dsRNA substrates (Supplement 5) suggesting that these structurally similar proteins may all share similar biochemical functions. It has been demonstrated that exogenous dsRNAs can affect gene function in early mouse embryos[29], and our results suggest that this regulation may be accomplished by an evolutionarily conserved RNAi machinery.

[0204] In addition to RNaseIII and helicase motifs, searches of the PFAM database indicate that each Dicer family member also contains a ZAP domain (**FIG. 8c**)[27]. This sequence was defined based solely upon its conservation in the Zwille/ARGONAUTE/Piwi family that has been implicated in RNAi by mutations in *C. elegans* (Rde-1)[9] and Neurospora (Qde-2)[10]. Although the function of this domain is unknown, it is intriguing that this region of homology is restricted to two gene families that participate in dsRNA-dependent silencing. Both the ARGONAUTE and Dicer families have also been implicated in common biological processes, namely the determination of stem-cell fates. A hypomorphic allele of carpel factory, a member of the Dicer family in Arabidopsis, is characterized by increased proliferation in floral meristems[24]. This phenotype and a number of other characteristic features are also shared by Arabidopsis ARGONAUTE (ago1-1) mutants[26] (C. Kidner and R. Martiennsen, pers. comm.). These genetic analyses begin to provide evidence that RNAi may be more than a defensive response to unusual RNAs but may also play important roles in the regulation of endogenous genes.

[0205] With the identification of Dicer as a catalyst of the initiation step of RNAi, we have begun to unravel the biochemical basis of this unusual mechanism of gene regulation. It will be of critical importance to determine whether the conserved family members from other organisms, particularly mammals, also play a role in dsRNA-mediated gene regulation.

[0206] Methods

[0207] Plasmid constructs. A full-length cDNA encoding Drosha was obtained by PCR from an EST sequenced by the Berkeley Drosophila genome project. The Homeless clone was a gift from Gillespie and Berg (Univ. Washington). The T7 epitope-tag was added to the amino terminus of each by PCR, and the tagged cDNAs were cloned into pRIP, a retroviral vector designed specifically for expression in insect cells (E. Bernstein, unpublished). In this vector, expression is driven by the *Orgyia pseudotsugata* IE2 promoter (Invitrogen). Since no cDNA was available for CG4792/Dicer, a genomic clone was amplified from a bacmid (BACR23F10; obtained from the BACPAC Resource Center in the Dept. of Human Genetics at the Roswell Park Cancer Institute). Again, during amplification, a T7 epitope tag was added at the amino terminus of the coding sequence. The human Dicer gene was isolated from a cDNA library prepared from HaCaT cells (GJH, unpublished). A T7-tagged version of the complete coding sequence was cloned into pCDNA3 (Invitrogen) for expression in human cells (LinX-A).

[0208] Cell culture and extract preparation. S2 and embryo culture. S2 cells were cultured at 27° C. in 5% CO_2 in Schneider's insect media supplemented with 10% heat inactivated fetal bovine serum (Gemini) and 1% antibiotic-antimycotic solution (Gibco BRL). Cells were harvested for extract preparation at 10×10^6 cells/ml. The cells were washed 1× in PBS and were resuspended in a hypotonic

buffer (10 mM Hepes pH 7.0, 2 mM MgCl2, 6 mM βME) and dounced. Cell lysates were spun 20,000xg for 20 minutes. Extracts were stored at −80° C. Drosophila embryos were reared in fly cages by standard methodologies and were collected every 12 hours. The embryos were dechorionated in 50% chlorox bleach and washed thoroughly with distilled water. Lysis buffer (10 mM Hepes, 10 mM KCl, 1.5 mM $MgCl_2$, 0.5 mM EGTA, 10 mM β-glycerophosphate, 1 mM DTT, 0.2 mM PMSF) was added to the embryos, and extracts were prepared by homogenization in a tissue grinder. Lysates were spun for two hours at 200,000xg and were frozen at −80° C. LinX-A cells, a highly-transfectable derivative of human 293 cells, (Lin Xie and GJH, unpublished) were maintained in DMEM/10%FCS.

[0209] Transfections and immunoprecipitations. S2 cells were transfected using a calcium phosphate procedure essentially as previously described[6]. Transfection rates were ~90% as monitored in controls using an in situ β-galactosidase assay. LinX-A cells were also transfected by calcium phosphate co-precipitation. For immunoprecipitations, cells (~$5x10^6$ per IP) were transfected with various clones and lysed three days later in IP buffer (125 mM KOAc, 1 mM MgOAc, 1 mM $CaCl_2$, 5 mM EGTA, 20 mM Hepes pH 7.0, 1 mM DTT, 1% NP-40 plus Complete protease inhibitors (Roche)). Lysates were spun for 10 minutes at 14,000xg and supernatants were added to T7 antibody-agarose beads (Novagen). Antibody binding proceeded for 4 hours at 4° C. Beads were centrifuged and washed in lysis buffer three times, and once in reaction buffer. The Dicer antiserum was raised in rabbits using a KLH-conjugated peptide corresponding to the C-terminal 8 amino acids of Drosophila Dicer-1 (CG4792).

[0210] Cleavage reactions. RNA preparation. Templates to be transcribed into dsRNA were generated by PCR with forward and reverse primers, each containing a T7 promoter sequence. RNAs were produced using Riboprobe (Promega) kits and were uniformly labeling during the transcription reaction with ^{32}P-UTP. Single-stranded RNAs were purified from 1% agarose gels. dsRNA cleavage. Five microliters of embryo or S2 extracts were incubated for one hour at 30° C. with dsRNA in a reaction containing 20 mM Hepes pH 7.0, 2 mM MgOAc, 2 mM DTT, 1 mM ATP and 5% Superasin (Ambion). Immunoprecipitates were treated similarly except that a minimal volume of reaction buffer (including ATP and Superasin) and dsRNA were added to beads that had been washed in reaction buffer (see above). For ATP depletion, Drosophila embryo extracts were incubated for 20 minutes at 30° C. with 2 mM glucose and 0.375 U of hexokinase (Roche) prior to the addition of dsRNA.

[0211] Northern and Western analysis. Total RNA was prepared from Drosophila embryos (0-12 hour), from adult flies, and from S2 cells using Trizol (Lifetech). Messenger RNA was isolated by affinity selection using magnetic oligo-dT beads (Dynal). RNAs were electrophoresed on denaturing formaldehyde/agarose gels, blotted and probed with randomly primed DNAs corresponding to Dicer. For Western analysis, T7-tagged proteins were immunoprecipitated from whole cell lysates in IP buffer using anti-T7-antibody-agarose conjugates. Proteins were released from the beads by boiling in Laemmli buffer and were separated by electrophoresis on 8% SDS PAGE. Following transfer to nitrocellulose, proteins were visualized using an HRP-conjugated anti-T7 antibody (Novagen) and chemiluminescent detection (Supersignal, Pierce).

[0212] RNAi of Dicer. Drosophila S2 cells were transfected either with a dsRNA corresponding to mouse caspase 9 or with a mixture of two dsRNAs corresponding to Drosophila Dicer-1 and Dicer-2 (CG4792 and CG6493). Two days after the initial transfection, cells were again transfected with a mixture containing a GFP expression plasmid and either luciferase dsRNA or GFP dsRNA as previously described[6]. Cells were assayed for Dicer activity or fluorescence three days after the second transfection. Quantification of fluorescent cells was done on a Coulter EPICS cell sorter after fixation. Control transfections indicated that Dicer activity was not affected by the introduction of caspase 9 dsRNA.

References Cited Example 2

[0213] 1. Baulcombe, D. C. RNA as a target and an initiator of post-transcriptional gene silencing in transgenic plants. *Plant Mol Biol* 32, 79-88 (1996).

[0214] 2. Wassenegger, M. & Pelissier, T. A model for RNA-mediated gene silencing in higher plants. *Plant Mol Biol* 37, 349-62 (1998).

[0215] 3. Montgomery, M. K. & Fire, A. Double-stranded RNA as a mediator in sequence-specific genetic silencing and co-suppression [see comments]. *Trends Genet* 14, 255-8 (1998).

[0216] 4. Sharp, P. A. RNAi and double-strand RNA. *Genes Dev* 13, 139-41 (1999).

[0217] 5. Sijen, T. & Kooter, J. M. Post-transcriptional gene-silencing: RNAs on the attack or on the defense? [In Process Citation]. *Bioessays* 22, 520-31 (2000).

[0218] 6. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* 404, 293-6 (2000).

[0219] 7. Hamilton, A. J. & Baulcombe, D. C. A species of small antisense RNA in posttranscriptional gene silencing in plants [see comments]. *Science* 286, 950-2 (1999).

[0220] 8. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* 101, 25-33 (2000).

[0221] 9. Tabara, H. et al. The rde-1 gene, RNA interference, and transposon silencing in *C. elegans*. *Cell* 99, 123-32 (1999).

[0222] 10. Catalanotto, C., Azzalin, G., Macino, G. & Cogoni, C. Gene silencing in worms and fungi. *Nature* 404, 245 (2000).

[0223] 11. Ketting, R. F., Haverkamp, T. H., van Luenen, H. G. & Plasterk, R. H. Mut-7 of *C. elegans*, required for transposon silencing and RNA interference, is a homolog of Werner syndrome helicase and RNaseD. *Cell* 99, 133-41 (1999).

[0224] 12. Cogoni, C. & Macino, G. Posttranscriptional gene silencing in Neurospora by a RecQ DNA helicase. *Science* 286, 2342-4 (1999).

[0225] 13. Cogoni, C. & Macino, G. Gene silencing in Neurospora crassa requires a protein homologous to RNA-dependent RNA polymerase. *Nature* 399, 166-9 (1999).

[0226] 14. Smardon, A. et al. EGO-1 is related to RNA-directed RNA polymerase and functions in germline development and RNA interference in *C. elegans* [published erratum appears in Curr Biol May 18, 2000;10(10):R393-4]. *Curr Biol* 10, 169-78 (2000).

[0227] 15. Mourrain, P. et al. Arabidopsis SGS2 and SGS3 genes are required for posttranscriptional gene silencing and natural virus resistance. *Cell* 101, 533-42 (2000).

[0228] 16. Dalmay, T., Hamilton, A., Rudd, S., Angell, S. & Baulcombe, D. C. An RNA-dependent RNA polymerase gene in Arabidopsis is required for post-transcriptional gene silencing mediated by a transgene but not by a virus. *Cell* 101, 543-53 (2000).

[0229] 17. Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P. & Sharp, P. A. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev* 13, 3191-7 (1999).

[0230] 18. Nicholson, A. W. Function, mechanism and regulation of bacterial ribonucleases. *FEMS Microbiol Rev* 23, 371-90 (1999).

[0231] 19. Filippov, V., Solovyev, V., Filippova, M. & Gill, S. S. A novel type of RNase III family proteins in eukaryotes. *Gene* 245, 213-21 (2000).

[0232] 20. Bass, B. L. Double-stranded RNA as a template for gene silencing. *Cell* 101, 235-8 (2000).

[0233] 21. Gillespie, D. E. & Berg, C. A. Homeless is required for RNA localization in Drosophila oogenesis and encodes a new member of the DE-H family of RNA-dependent ATPases. *Genes Dev* 9, 2495-508 (1995).

[0234] 22. Robertson, H. D., Webster, R. E. & Zinder, N. D. Purification and properties of ribonuclease III from *Escherichia coli*. *J Biol Chem* 243, 82-91 (1968).

[0235] 23. Dunn, J. J. RNase III cleavage of single-stranded RNA. Effect of ionic strength on the fidelity of cleavage. *J Biol Chem* 251, 3807-14 (1976).

[0236] 24. Jacobsen, S. E., Running, M. P. & Meyerowitz, E. M. Disruption of an RNA helicase/RNAse III gene in Arabidopsis causes unregulated cell division in floral meristems. *Development* 126, 5231-43 (1999).

[0237] 25. Matsuda, S. et al. Molecular cloning and characterization of a novel human gene (HERNA) which encodes a putative RNA-helicase. *Biochim Biophys Acta* 1490, 163-9 (2000).

[0238] 26. Bohmert, K. et al AGO1 defines a novel locus of Arabidopsis controlling leaf development. *Embo J* 17, 170-80 (1998).

[0239] 27. Sonnhammer, E. L., Eddy, S. R. & Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**, 405-20 (1997).

[0240] 28. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402 (1997).

[0241] 29. Wianny, F. and Zernicka-Goetz, M. Specific interference with gene function by double-stranded RNA in early mouse development. *Nature Cell Biol.* 2, 70-75 (2000).

[0242] 30. Fagard, M., Boutet, S., Morel, J.-B., Bellini, C. and Vaucheret, H. Ago-1, Qde-2 and Rde-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc. Natl. Acad. Sci. USA* 97, 11650-11654 (2000).

Example 3

A simplified Method for the Creation of Hairpin Constructs for RNA Interference

[0243] In numerous model organisms, double stranded RNAs have been shown to cause effective and specific suppression of gene function (ref. 1). This response, termed RNA interference or post-transcriptional gene silencing, has evolved into a highly effective reverse genetic tool in *C. elegans*, Drosophila, plants and numerous other systems. In these cases, double-stranded RNAs can be introduced by injection, transfection or feeding; however, in all cases, the response is both transient and systemic. Recently, stable interference with gene expression has been achieved by expression of RNAs that form snap-back or hairpin structures (refs 2-7). This has the potential not only to allow stable silencing of gene expression but also inducible silencing as has been observed in trypanosomes and adult Drosophila (refs 2,4,5). The utility of this approach is somewhat hampered by the difficulties that arise in the construction of bacterial plasmids containing the long inverted repeats that are necessary to provoke silencing. In a recent report, it was stated that more than 1,000 putative clones were screed to identify the desired construct (ref 7).

[0244] The presence of hairpin structures often induces plasmid rearrangement, in part due to the *E. coli* nuc proteins that recognize and cleave cruciform DNA structures (ref 8). We have developed a method for the construction of hairpins that does not require cloning of inverted repeats, per se. Instead, the fragment of the gene that is to be silenced is cloned as a direct repeat, and the inversion is accomplished by treatment with a site-specific recombinase, either in vitro (or potentially in vivo) (see **FIG. 27**). Following recombination, the inverted repeat structure is stable in a bacterial strain that lacks an intact SBC system (DL759). We have successfully used this strategy to construct numerous hairpin expression constructs that have been successfully used to provoke gene silencing in Drosophila cells.

Literature Cited in Example 3

[0245] 1. Bosher, J. M. & Labouesse, M. RNA interference: genetic wand and genetic watchdog. *Nat Cell Biol* 2, E31-6 (2000).

[0246] 2. Fortier, E. & Belote, J. M. Temperature-dependent gene silencing by an expressed inverted repeat in Drosophila [published erratum appears in Genesis; May 27, 2000; (1):47]. *Genesis* 26, 240-4 (2000).

[0247] 3. Kennerdell, J. R. & Carthew, R. W. Heritable gene silencing in Drosopbila using double-stranded RNA. *Nat Biotechnol* 18, 896-8 (2000).

[0248] 4. Lam, G. & Thummel, C. S. Inducible expression of double-stranded RNA directs specific genetic interference in Drosopbila [In Process Citation]. *Curr Biol* 10, 957-63 (2000).

[0249] 5. Shi, H. et al. Genetic interference in *Trypanosoma brucei* by heritable and inducible double-stranded RNA. *Rna* 6, 1069-76 (2000).

[0250] 6. Smith, N. A. et al Total silencing by intron-spliced bairpin RNAs. *Nature* 407, 319-20 (2000).

[0251] 7. Tavernarakis, N., Wang, S. L., Dorovkov, M., Ryazanov, A. & Driscoll, M. Heritable and inducible genetic interference by double-stranded RNA encoded by transgenes. *Nat Genet* 24, 180-3 (2000).

[0252] 8. Connelly, J. C. & Leacb, D. R. The sbcC and sbcD genes of *Escherichia coli* encode a nuclease involved in palindrome inviability and genetic recombination. *Genes Cells* 1, 285-91 (1996).

V. EQUIVALENTS

[0253] Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, many equivalents to the specific embodiments of the invention described herein. Such equivalents are intended to be encompassed by the following claims.

[0254] All of the above-cited references and publications are bereby incorporated by reference.

We claim:

1. A method for attenuating expression of a target gene in cultured cells, comprising introducing double stranded RNA (dsRNA) into the cells in an amount sufficient to attenuate expression of the target gene, wherein the dsRNA comprises a nucleotide sequence that hybridizes under stringent conditions to a nucleotide sequence of the target gene.

2. A metbod for attenuating expression of a target gene in a mammalian cell, comprising

(i) activating one or both of a Dicer activity or an Argonaut activity in the cell, and

(ii) introducing into tbe cell a double stranded RNA (dsRNA) in an amount sufficient to attenuate expression of tbe target gene, wberein the dsRNA comprises a nucleotide sequence that hybridizes under stringent conditions to a nucleotide sequence of tbe target gene.

3. The method of claim 2, wherein the cell is suspended in culture.

4. The method of claim 2, wherein the cell is in a whole animal, such as a non-human mammal.

5. The method of claim 1 or **2**, wberein is engineered with (i) a recombinant gene encoding a Dicer activity, (ii) a recombinant gene encoding an Argonaut activity, or (iii) both.

6. The metbod of claim 5, wberein the recombinant gene encodes a protein whicb includes an amino acid sequence at least 50 percent identical to SEQ ID No. **2** or 4 or the Argonaut sequence shown in **FIG. 24**.

7. The metbod of claim 5, wberein the recombinant gene includes a coding sequence bybridizes under wash conditions of 2xSSC at 22° C. to SEQ 1D No. 1 or 3.

8. The method of claim 1 or **2**, wherein an endogenous Dicer gene or Argonaut gene is activated.

9. The method of claim 1 or **2**, wherein the target gene is an endogenous gene of the cell.

10. The method of claim 1 or **2**, wherein the target gene is an heterologous gene relative to the genome of the cell, such as a pathogen gene.

11. The method of claim 1 or **2**, wherein the cell is treated with an agent that inbibits protein kinase RNA-activated (PKR) apoptosis, such as by treatment with agents whicb inbibit expression of PKR, cause its destruction, and/or inhibit the kinase activity of PKF.

12. The method of claim 1 or **2**, wherein the cell is a primate cell, such as a human cell.

13. The method of claim 1 or **2**, wherein the dsRNA is at least 20 nucleotides in length.

14. The method of claim 13, wherein the dsRNA is at least 100 nucleotides in length.

15. The metbod of claim 1 or **2**, wberein expression of tbe target gene is attenuated by at least 10 fold.

16. An assay for identifying nucleic acid sequences responsible for conferring a particular pbenotype in a cell, comprising

(i) constructing a variegated library of nucleic acid sequences from a cell in an orientation relative to a promoter to produce double stranded DNA;

(ii) introducing the variegated dsRNA library into a culture of target cells, which cells have an activated Dicer activity or Argonaut activity;

(iii) identifying members of the library which confer a particular phenotype on the cell, and identifying the sequence from a cell which correspond, such as being identical or bomologous, to the library member.

17. A method of conducting a drug discovery business comprising:

(i) identifying, by the assay of claim 16, a target gene which provides a pbenotypically desirable response when inhibited by RNAi;

(ii) identifying agents by their ability to inhibit expression of the target gene or the activity of an expression product of tbe target gene;

(iii) conducting therapeutic profiling of agents identified in step (b), or further analogs thereof, for efficacy and toxicity in animals; and

(iv) formulating a pharmaceutical preparation including one or more agents identified in step (iii) as having an acceptable therapeutic profile.

18. The method of claim 17, including an additional step of establishing a distribution system for distributing tbe pbarmaceutical preparation for sale, and may optionally include establishing a sales group for marketing the pharmaceutical preparation.

19. A method of conducting a target discovery business comprising:

(i) identifying, by the assay of claim 16, a target gene which provides a phenotypically desirable response when inhibited by RNAi;

Intron-mediated recombinant techniques and reagents

Abstract

The present invention makes available methods and reagents for novel manipulation of nucleic acids. As described herein, the present invention makes use of the ability of intronic sequences, such as derived from group I, group II, or nuclear pre-mRNA introns, to mediate specific cleavage and ligation of discontinuous nucleic acid molecules. For example, novel genes and gene products can be generated by admixing nucleic acid constructs which comprise exon nucleic acid sequences flanked by intron sequences that can direct trans-splicing of the exon sequences to each other. The flanking intronic sequences can, by intermolecular complementation, form a reactive complex which promotes the transesterification reactions necessary to cause the ligation of discontinuous nucleic acid sequences to one another, and thereby generate a recombinant gene comprising the ligated exons.

RELATED APPLICATIONS

This application is a continuation-in-part of U.S. Ser. No. 08/119,512, filed Sep. 10, 1993, now U.S. Pat. No. 5,498,531, entitled "Intron-Mediated Recombinant Techniques and Reagents", the specification of which is incorporated by reference herein.

Claims

I claim:

1. A purified preparation of a reverse-splicing intron, which reverse-splicing intron comprises:

a first segment comprising a 5' portion of a group II intron, which 5' portion includes an exon binding site not naturally present in said group II intron; and

a second segment comprising a 3' portion of a group II intron, which 3' portion includes a domain V motif, a branch site acceptor forming a phosphodiester bond with a 5' end of said first segment, and a nucleophilic group at a 3' end of said second segment for transesterifying a phosphodiester bond of a ribonucleic acid,

wherein said first and second segments together form an autocatalytic y-branched intron which catalyzes integration of at least the first segment of the reverse-splicing intron into a substrate ribonucleic acid by a reverse-splicing reaction.

2. The reverse-splicing intron of claim 1, wherein said 5' portion of the group II intron comprises intron domains V and VI, and said 3' portion of the group II intron comprises intron domains I-III.

3. The reverse-splicing intron of claim 2, which reverse-splicing intron is represented by the general formula:

where

(IVS 1-3) represents a 5' portion of a group II intron, (IVS5,6) represents a 3' portion of a group II intron, which portion includes a branch site acceptor, and

2'-5' represents a phosphodiester bond formed between a branch site acceptor of (IVS5,6) and the 5' end of (IVS1-3), (IVS1-3) and (IVS5,6) otherwise being discontinuous with each other,

wherein (IVS 1-3) and (IVS5,6) together form an autocatalytic Y-branched intron which catalyzes integration of the (IVS1-3) fragment into a substrate ribonucleic acid by a reverse-splicing reaction.

4. The reverse-splicing intron of claim 1, wherein said first and second segments are contiguous, via a covalent bond other than the phosphodiester bond formed with said branch site acceptor, and form a y-branched lariat.

5. The reverse-splicing intron of claims 2 or 4, which reverse-splicing intron is represented by the general formula: ##STR2## wherein (IVS1-3) represents a 5' portion of a group II intron,

(IVS5,6) represents a 3' portion of a group II intron, which portion includes a branch site acceptor,

'- ' represents a phosphodiester bond formed between a branch site acceptor of (IVS5,6) and the 5' end of (IVS1-3), and

A represents a phosphodiester bond between a 3' end of (IVS1-3) and a 5' end of (IVS5,6),

wherein (IVS1-3) and (IVS5,6) together form an autocatalytic Y-branched lariat which catalyzes integration of the reverse-splicing intron into a substrate ribonucleic acid by a reverse-splicing reaction.

6. The reverse-splicing intron of claims 1, 2 or 4, wherein said exon binding site is selected to provide specific integration into the substrate ribonucleic acid after a selected intron binding site, which intron binding site is from 3-16 nucleotides in length.

7. The reverse-splicing intron of claim 6, wherein said intron binding site is from 5-7 nucleotides in length.

8. The reverse-splicing intron of claim 1, wherein said exon binding site comprises an EBS1 and an EBS2.

9. A purified preparation of a reverse-splicing construct, which reverse-splicing construct comprises two or more fragments of autocatalytic introns and catalyzes integration of at least a portion of the reverse-splicing construct into a substrate ribonucleic acid by a reverse-splicing reaction, wherein the portion of the intron sequences that provides sequence specificity for the

substrate ribonucleic acid differs in structure and specificity from the naturally-occurring sequence of the autocatalytic intron.

10. The preparation of claim 9, wherein the autocatalytic intron fragments are selected from group II introns.

11. The preparation of claim 10, wherein the reverse-splicing construct comprises a 5' portion of a group II intron including intron domains V and VI, and a 3' portion of a group II intron including intron domains I-III.

12. The preparation of claim 10, wherein the reverse-splicing construct comprises, from a group I intron, an internal guide sequence, a GTP-binding site, wherein said group I intron fragments reconstitute a functional intron through intermolecular complementation.

13. The preparation of claim 9, wherein the autocatalytic intron fragments are selected from group I introns.

14. A library of reverse-splicing introns comprising a variegated population of purified y-branched group II introns, which variegated population is characterized as including at least 25 different y-branched group II introns of unique specificity.

15. The library of reverse-splicing introns of claim 14, wherein the variegated population includes at least 100 different y-branched group II introns of unique specificity.

16. The library of reverse-splicing introns of claim 14, wherein the variegated population includes from 10.sup.3 to 10.sup.5 different y-branched group II introns of unique specificity.

17. The library of reverse-splicing introns of claim 14, wherein the y-branched group II introns include:

a first segment comprising a 5' portion of a group II intron, which 5' portion includes an exon binding site not naturally present in said group II intron; and

a second segment comprising a 3' portion of a group II intron, which 3' portion includes a domain V motif, a branch site acceptor forming a phosphodiester bond with a 5' end of said first segment, and a nucleophilic group at a 3' end of said second segment for transesterifying a phosphodiester bond of a ribonucleic acid,

wherein said first and second segments together form an autocatalytic y-branched intron which catalyzes integration of at least the first segment of the reverse-splicing intron into a substrate ribonucleic acid by a reverse-splicing reaction.

18. The library of reverse-splicing introns of claim 17, wherein said 5' portion of the group II intron comprises intron domains V and VI, and said 3' portion of the group II intron comprises intron domains I-III.

19. The library of reverse-splicing intron of claim 17, wherein said first and second segments are contiguous, via a covalent bond other than the phosphodiester bond formed with said branch site acceptor, and form a y-branched lariat.

20. The library of reverse-splicing intron of claims 14, 18 or 19, which reverse-splicing intron is represented by the general formula: ##STR3## wherein (IVS1-3) represents a 5' portion of a group II intron,

(IVS5,6) represents a 3' portion of a group II intron, which portion includes a branch site acceptor,

'- ' represents a phosphodiester bond formed between a branch site acceptor of (IVS5,6) and the 5' end of(IVS1-3), and

A, if present, represents a phosphodiester bond between a 3' end of (IVS1-3) and a 5' end of (IVS5,6),

wherein (IVS1-3) and (IVS5,6) together form an autocatalytic Y-branched intron which catalyzes integration of the reverse-splicing intron into a substrate ribonucleic acid by a reverse-splicing reaction.

21. A method for generating a chimeric ribonucleic acid by trans-splicing, comprising admixing two or more splicing constructs under trans-splicing reaction conditions, which splicing constructs comprise a ribonucleic acid represented by the general formula (3' IVS)-EX-(5' IVS), wherein

EX represents an exonic ribonucleic acid sequence which is intended to be present in a chimeric ribonucleic acid, said exonic sequence having a 5' exon end and a 3' exon end,

(3' IVS) is absent or represents a 3' fragment of an intron, which 3' intron fragment is covalently attached to the 5' exon end of said exonic sequence by a phosphodiester bond, and

(5' IVS) is absent or represents a 5' fragment of an intron, which 5' intron fragment is covalently attached to the 3' exon end of said exonic sequence by a phosphodiester bond,

with the proviso that

at least one of (3' IVS) and (5' IVS) is present on each splicing construct; and

the intron fragments for at least a portion of said splicing constructs are not from intronic sequences which naturally flank the corresponding covalently attached exonic sequence;

wherein said exons of said set of splicing constructs comprise a variegated population of ribonucleic acids, and said trans-splicing reaction conditions comprise conditions in which 3' and 5' intron fragments of different splicing constructs reconstitute a functional intron through intermolecular complementation and ligate said exons to generate said chimeric ribonucleic acid.

22. The method of claim 21, wherein at least a portion of said exonic sequences are spliced to each other in predetermined order.

23. The method of claim 21, wherein said 3' and 5' intron fragments of said splicing constructs comprise group II intron fragments including,

i) an exon binding site, and

ii) a branch site acceptor comprising an activated nucleophile for forming a phosphodiester bond with a 5' intron end of said 5' intron fragment and for cleaving said 5' intron fragment from the 3' end of said exonic sequence.

24. The method of claim 23, wherein said group II intron fragments further comprise at least a portion of a group II domain V sufficient to reconstitute said functional intron.

25. The method of claim 23, wherein said trans-splicing reaction conditions further comprises admixing with said splicing constructs at least a portion of a domain V of a group II intron sufficient to interact with said 3' and 5' intron fragments and reconstitute said functional intron.

26. The method of claim 21, wherein said 3' and 5' intron fragments of said splicing constructs comprise group I intron fragments including an internal guide sequence, a GTP-binding site, and a 3' terminal G located in said 3' intron fragment immediately adjacent said 5' exon end of said exonic sequence.

27. The method of claim 21, wherein said 3' and 5' intron fragments of said splicing constructs comprise nuclear pre-mRNA intron fragments including a 5' splice junction sequence, a 3' splice junction sequence, and a branchpoint sequence; and

said trans-splicing reaction conditions include admixing, with said splicing constructs, adenosine triphosphate (ATP) and small nuclear ribonucleoproteins (snRNPs).

28. The method of claim 27, wherein said snRNPs comprise a U1 snRNP, a U2 snRNP, a U4 snRNP, a U5 snRNP, and a U6 snRNP.

29. The method of claim 21, wherein said exonic sequence comprises a polypeptide encoding sequence.

30. The nucleic acid construct of claim 29, wherein said 3' and 5' intron fragments comprise nuclear pre-mRNA intron fragments.

31. The method of claim 21, wherein said exonic sequence comprises a ribozyme sequence.

32. A method for generating a chimeric ribonucleic acid by trans-splicing, comprising admixing a two or more splicing constructs under trans-splicing reaction conditions, which splicing constructs comprise a ribonucleic acid represented by the general formula (3' IVS)-EX-(5' IVS),

wherein

EX represents an ribonucleic acid encoding at least a portion of a mammalian polypetide, said exon having a 5' exon end and a 3' exon end,

(3' IVS) is absent or represents a 3' fragment of an intron, which 3' intron fragment is covalently attached to the 5' exon end of said exon by a phosphodiester bond, and

(5' IVS) is absent or represents a 5' fragment of an intron, which 5' intron fragment is covalently attached to the 3' exon end of said exon by a phosphodiester bond,

with the proviso that at least one of (3' IVS) and (5' IVS) is present on each splicing construct,

wherein said exons of said set of splicing constructs comprise a variegated population of polypeptide-encoding sequences, and said trans-splicing reaction conditions comprise conditions in which 3' and 5' intron fragments of different splicing constructs reconstitute a functional intron through intermolecular complementation and ligate said exons to generate said chimeric ribonucleic acid.

33. A library of splicing constructs comprising a variegated population of nucleic acids each represented by the general formula (3' IVS)-EX-(5' IVS), wherein

EX represents an exonic ribonucleic acid which is intended to be present in a chimeric ribonucleic acid, said exon having a 5' exon end and a 3' exon end,

(3' IVS) is absent or represents a 3' fragment of an intron, which 3' intron fragment is covalently attached to the 5' exon end of said exon by a phosphodiester bond, and

(5' IVS) is absent or represents a 5' fragment of an intron, which 5' intron fragment is covalently attached to the 3' exon end of said exon by a phosphodiester bond,

with the proviso that

at least one of (3' IVS) and (5' IVS) is present on each splicing construct; and

the intron fragments for at least a portion of said splicing constructs are not from intronic sequences which naturally flank the corresponding covalently attached exon;

wherein said exons of said set of splicing constructs comprise a variegated population of ribonucleic acids, and said trans-splicing reaction conditions comprise conditions in which 3' and 5' intron fragments of different splicing constructs reconstitute a functional intron through intermolecular complementation and ligate said exons to generate said chimeric ribonucleic acid.

34. A splicing construct comprising a nucleic acid represented by the general formula (3' IVS)-EX-(5' IVS), wherein

EX represents an ribonucleic acid encoding at least a portion of a mammalian polypetide, said exon having a 5' exon end and a 3' exon end,

(3' IVS) is absent or represents a 3' fragment of an intron, which 3' intron fragment is covalently attached to the 5' exon end of said exon by a phosphodiester bond, and

(5' IVS) is absent or represents a 5' fragment of an intron, which 5' intron fragment is covalently attached to the 3' exon end of said exon by a phosphodiester bond,

with the proviso that

at least one of (3' IVS) and (5' IVS) is present on each splicing construct,

wherein said exon sequence is discontinuous with any nucleic acid sequences other than said flanking intron sequences, and said 3' and 5' intron fragments can, through intermolecular complementation, mediate trans-splicing reactions between two or more of said splicing constructs.

35. The nucleic acid construct of claim 34, wherein said 3' and 5' intron fragments comprise group II intron fragments.

36. The nucleic acid construct of claim 34, wherein said 3' and 5' intron fragments comprise group I intron fragments.

37. A method for generating a circular ribonucleic acid by intron-mediated splicing comprising, providing a ribonucleic acid, represented by the general formula (3' IVS)-EX-(5' IVS), wherein

EX represents an exonic nucleic acid which is intended to be present in a circular nucleic acid, said exon having a 5' exon end and a 3' exon end,

(3' IVS) represents a 3' fragment of an intron, which 3' intron fragment is covalently attached to the 5' exon end of said exon by a phosphodiester bond, and

(5' IVS) represents a 5' fragment of an intron, which 5' intron fragment is covalently attached to the 3' exon end of said exon by a phosphodiester bond, under trans-splicing conditions which cause reconstitution of a functional intron by intramolecular complementation of said 3' and 5' intron fragments, which functional intron, under said trans-splicing conditions, ligatates said 3' and 5' end of said exon to generate a circular ribonucleic acid.

38. The method of claim 37, wherein said trans-splicing conditions include providing a bridging oligonucleotide with said ribonucleic acid to facilitate intramolecular complementation of said 3' and 5' intron fragments.

39. The method of claim 37, wherein the ribonucleic acid includes a structural element which non-covalently links the (3' IVS) and (5' IVS).

40. The method of claim 39, wherein (3' IVS) and (5' IVS) each contain complementary sequences which anneal to non-covalently link the (3' IVS) and (5' IVS).

41. The method of claim 37, wherein at least one of the (3' IVS) or (5' IVS) represents a fragment of an intron not normally associated with the exonic nucleic acid.

42. The method of claims 37, wherein (3' IVS) or (5' IVS) represent fragments of an autocatalytic intron, and combine through intramolecular complementation to form a functional intron.

43. The method of claim 42, wherein (3' IVS) or (5' IVS) represent fragments of a group II autocatalytic intron.

44. The method of any of claims 37, 39-43, wherein the ribonucleic acid is produced by transcription of an expression construct in a host cell, and the trans-splicing which generates the circular ribonucleic acid occurs in the host cell.

45. The method of claim 44, wherein the host cell is a mammalian cell.

Description

BACKGROUND OF THE INVENTION

Most eukaryotic genes are discontinuous with proteins encoded by them, consisting of coding sequences (exons) interrupted by non-coding sequences (introns). After transcription into RNA, the introns are removed by splicing to generate the mature messenger RNA (mRNA). The splice points between exons are typically determined by consensus sequences that act as signals for the splicing process.

Structural features of introns and the underlying splicing mechanisms form the basis for classification of different kinds of introns. Since RNA splicing was first described, four major categories of introns have been recognized. Splicing of group I, group II, nuclear pre-mRNA, and tRNA introns can be differentiated mechanistically, with certain group I and group II introns able to be autocatalytically excised from a pre-RNA in vitro in the absence of any other protein or RNA factors. In the instance of the group I, group II and nuclear pre-mRNA introns, splicing proceeds by a two-step transesterification mechanism.

To illustrate, the nuclear rRNA genes of certain lower eukaryotes (e.g., Tetrahymena thermophila and Physarum polycephalum) contain group I introns. This type of intron also occurs in chloroplast, yeast, and fungal mitochondrial rRNA genes; in certain yeast and fungal mitochondrial mRNA; and in several chloroplast tRNA genes in higher plants. Group I introns are characterized by a linear array of conserved sequences and structural features, and are excised by two successive transesterifications. Splicing of the Tetrahymena pre-rRNA intron, a prototypic group I intron, proceeds by two transesterification reactions during which phosphate

esters are exchanged without intermediary hydrolysis. Except for the initiation step, promoted by a free guanosine, all reactive groups involved in the transesterification reactions are contained within the intron sequence The reaction is initiated by the binding of guanosine to an intron sequence. The unshared pair of electrons of the 3'-hydroxyl group of the bound guanosine can act as a nucleophile, attacking the phosphate group at the 5' exon-intron junction (splice site), resulting in cleavage of the precursor RNA. A free 3'-hydroxyl group is generated at the cleavage site (the end of the 5' exon) and release of the intron occurs in a second step by attack of the 5' exon's 3'-hydroxyl group on the 3' splice site phosphate.

Group II introns, which are classed together on the basis of a conserved secondary structure, have been identified in certain organellar genes of lower eukaryotes and plants.

The group II introns also undergo self-splicing reactions in vitro, but in this instance, a residue within the intron, rather than added guanosine, initiates the reaction. Another key difference between group II and group I introns is in the structure of the excised introns. Rather than the linear products formed during splicing of group I introns, spliced group II introns typically occur as lariats, structures in which the 5'-phosphoryl end of the intron RNA is linked through a phosphodiester bond to the 2'-hydroxyl group of an internal nucleotide. As with group I introns, the splicing of group II introns occurs via two transesterification steps, one involving cleavage of the 5' splice site and the second resulting in cleavage of the 3' splice site and ligation of the two exons. For example, 5' splice site cleavage results from nucleophilic attack by the 2'-hydroxyl of an internal nucleotide (typically an adenosine) located upstream of the 3' splice site, causing the release of the 5' exon and the formation of a lariat intermediate (so called because of the branch structure of the 2', 5' phosphodiester bond thus produced). In the second step, the 3'-end hydroxyl of the upstream exon makes a nucleophilic attack on the 3' splice site. This displaces the intron and joins the two exons together.

Eukaryotic nuclear pre-mRNA introns and group II introns splice by the same mechanism; the intron is excised as a lariat structure, and the two flanking exons are joined. Moreover, the chemistry of the two processes is similar. In both, a 2'-hydroxyl group within the intron serves as the nucleophile to promote cleavage at the 5' splice site, and the 3' hydroxyl group of the upstream exon is the nucleophile that cleaves the 3' splice site by forming the exon-exon bond. However, in contrast to the conserved structural elements that reside within group I and II introns, the only conserved features of nuclear pre-mRNA introns are restricted to short regions at or near the splice junctions. In yeast, these motifs are (i) a conserved hexanucleotide at the 5' splice, (ii) an invariant heptanucleotide, the UACUAAC Box, surrounding the branch point A, (iii) a generally conserved enrichment for pyrimidine residues adjacent to the invariant AG dinucleotide at the 3' splice site. Further characteristics of nuclear pre-mRNA splicing in vitro that distinguish it from autocatalytic splicing are the dependence on added cell-free extracts, and the requirement for adenosine triphosphate (ATP) Another key difference is that nuclear pre-mRNA splicing generally requires multiple small nuclear ribonucleoproteins (snRNPs) and other accessory proteins, which can make-up a larger multi-subunit complex (splicesome) that facilitates splicing.

SUMMARY OF THE INVENTION

The present invention makes available methods and reagents for novel manipulation of nucleic acids. As described herein, the present invention makes use of the ability of intronic sequences, such as derived from group I, group II, group III or nuclear pre-mRNA introns, to mediate specific cleavage and ligation of discontinuous nucleic acid molecules. For example, novel genes and gene products can be generated by admixing nucleic acid constructs comprising "exon" nucleic acid sequences flanked by intron sequences that can direct transsplicing of the exon sequences to each other. The flanking intronic sequences, by intermolecular complementation between the flanking intron sequences of two different constructs, form a functional intron which mediates the transesterification reactions necessary to cause the ligation of the discontinuous nucleic acid sequences to one another, and thereby generate a recombinant gene comprising the ligated exons. As used herein, the term exon denotes nucleic acid sequences, or exon "modules", that can, for instance, encode portions of proteins or polypeptide chains, such as corresponding to naturally occurring exon sequences or naturally occurring exon sequences which have been mutated (e.g. point mutations, truncations, fusions), as well as nucleic acid sequences from "synthetic exons" including sequences of purely random construction. However, the term "exon", as used in the present invention, is not limited to protein-encoding sequences, and may comprises nucleic acid sequences of other function, including nucleic acids of "intronic origin" which give rise to, for example, ribozymes or other nucleic acid structure having some defined chemical function.

As described herein, novel genes and gene products can be generated, in one embodiment of the present method, by admixing nucleic acid constructs which comprise a variegated population of exon sequences. As used herein, variegated refers to the fact that the population includes nucleic acids of different nucleotide compositions. When the interactions of the flanking introns are random, the order and composition of the internal exons of the combinatorial gene library generated is also random. For instance, where the variegated population of exons used to generate the combinatorial genes comprises N different internal exons, random trans-splicing of the internal exons can result in N.sup.y different genes having y internal exons. However, the present trans-splicing method can also be utilized for ordered gene assembly such that nucleic acid sequences are spliced together in a predetermined order, and can be carried out in much the same fashion as automated oligoucleotide or polypeptide synthesis. In similar fashion, an ordered combinatorial ligation can be carried out in which particular types of exons are added to one and other in an ordered fashion, but, at certain exon positions, more than one type of exon may be added to generate a library of combinatorial genes.

Furthermore,the present invention makes available methods and reagents for producing circular RNA molecules. In particular, exon constructs flanked by either group II or nuclear pre-mRNA fragments can, under conditions which facilitate exon ligation by splicing of the flanking intron sequences, drive the manufacture of circularly permuted exonic sequences in which the 5' and 3' ends of the same exon are covalently linked via a phosphodiester bond. Circular RNA moieties generated in the present invention can have several advantages over the equivalent "linear" constructs. For example, the lack of a free 5' or 3' end may render the molecule less susceptible to degradation by cellular nucleases. Such a characteristic can be especially beneficial, for instance, in the use of ribozymes in vivo, as might be involved in a particular gene therapy. The circularization of mature messenger-RNA transcripts can also be beneficial, by conferring increased stability as described above, as well as potentially increasing the level of protein

translation from the transcript.

DESCRIPTION OF DRAWINGS

FIG. 1 is a schematic representation of the group II splicing reaction, as well as the reverse-slicing reaction.

FIG. 2 illustrates the domain structure of a group II intron.

FIG. 3 is a schematic representation of an illustrative group I splicing reaction, as well as a reverse-splicing reaction.

FIG. 4 illustrates the secondary structure of a group I intron.

FIG. 5 is a schematic representation of a trans-splicing reaction between discontinuous exon sequences.

FIG. 6 illustrates how a reverse-splicing reaction can be utilized to activate exons for subsequent combinatorial trans-splicing.

FIG. 7 illustrates an ordered gene assembly mediated by trans-splicing of exons flanked with nuclear pre-mRNA intron fragments.

FIG. 8 illustrates the consensus sequence for group IIA and IIB domain V.

FIG. 9A illustrates the interaction between nuclear pre-mRNA introns and snRNPs.

FIGS. 9B and 9C illustrate two embodiments for accomplishing nuclear pre-mRNA intron mediated trans-splicing.

FIG. 10 is a schematic representation of an intron-mediated combinatorial method which relies on cis-splicing to ultimately form the chimeric genes.

FIG. 11 depicts one example of how group I intron sequences can be used to shuffle group II intron domains.

FIG. 12 illustrates an "exon-trap" assay for identifying exons from genomic DNA, utilizing trans-splicing mediated by discontinuous nuclear pre-mRNA intron fragments.

FIG. 13A shows a nucleic acid construct, designated (IVS5,6)-exon-(IVS1-3), which can mediate trans-splicing between heterologous exons, as well as be used to generate circular RNA transcripts.

FIGS. 13B depicts two a nucleic acid construct, designated (3'-half-IVS)-exon-(5'-half-IVS), which can mediate trans-splicing between heterologous exons, as well as be used to generate circular RNA transcripts.

FIG. 14 shows how group II intronic fragments can be utilized to covalently join the ends of a nuclear pre-mRNA transcripts having flanking nuclear pre-mRNA intron fragments, such that the flanking nuclear pre-mRNA intron fragments can subsequently drive ligation of the 5' and 3' end of the exonic sequences.

FIGS. 15A-C illustrate how intronic ends of the same molecule can be brought together by a nucleic acid "bridge" which involves hydrogen bonding between the intronic fragments flanking an exon and a second discrete nucleic acid moiety.

FIG. 15D shows, in an illustrative embodiment, how a nucleic acid bridge can be used to direct alternative splicing by "exon skipping".

FIG. 16 illustrates a nucleic acid construct useful in mediating the alternate splicing of an exon through a trans-splicing-like mechanism.

FIG. 17 is an exemplary illustration of the generation of recombinant Y-branched group II lariats.

FIG. 18 depicts a further embodiment illustrating how a reverse-splicing ribozyme, such as the group II lariat IVS, can also be used to cleave and ligate target RNA molecules.

FIG. 19 depicts a method by which the present trans-splicing constructs can be used to manipulate nucleic acid sequences into a plasmid such as a cloning or expression vector.

FIG. 20A is an illustration of the composite protein structure of the variable region of both heavy and light chains of an antibody.

FIGS. 20B-C illustrate possible combinatorial constructs produced using antibody framework regions (FRs) and complementarity determining regions (CDRs).

DETAILED DESCRIPTION OF THE INVENTION

Biological selections and screens are powerful tools with which to probe protein and nucleic acid function and to isolate variant molecules having desirable properties. The technology described herein enables the rapid and efficient generation and selection of novel genes and gene products. The present combinatorial approach, for example, provides a means for capturing the vast diversity of exons, and relies on the ability of intron sequences to mediate random splicing between exons.

As described below, novel genes and gene products can be generated, in one embodiment of the present combinatorial method, by admixing a variegated population of exons which have flanking intron sequences that can direct trans-splicing of the exons to each other. Under conditions in which trans-splicing occurs between the exons, a plurality of genes encoding a combinatorial library are generated by virtue of the ability of the exons to be ligated together in a random fashion. Where the initial variegated exon population are ribonucleotides (i.e. RNA), the

resulting combinatorial transcript can be reverse-transcribed to cDNA and cloned into an appropriate expression vector for further manipulation or screening.

In another embodiment of the present combinatorial method, a variegated population of single-stranded DNA molecules corresponding to exon sequences of both (+) and (-) strand polarity, and which have flanking intron sequences capable of mediating cis-splicing, are provided together such that a portion of the nucleic acid sequence in the flanking intron of an exon of one polarity (e.g. a (+) strand) can base pair with a complementary sequence in the flanking intron of another exon of opposite polarity (e.g. a (-) strand). Using standard techniques, any single-stranded regions of the concatenated exon/intron sequences can be subsequently filled-in with a polymerase, and nicks covalently closed with a ligase, to form a double-stranded chimeric gene comprising multiple exons interrupted by intron sequences. Upon transcription of the chimeric gene to RNA, cis-splicing can occur between the exons of the chimeric gene to produce the mature RNA transcript, which can encode a chimeric protein.

As used herein, the term "exon" denotes nucleic acid sequences, or exon "modules", which intended to be retained in the gene created by the subject method. For instance, exons can encode portions of proteins or polypeptide chains The exons can correspond to discrete domains or motifs, as for example, functional domains, folding regions, or structural elements of a protein; or to short polypeptide sequences, such as reverse turns, loops, glycosylation signals and other signal sequences, or unstructured polypeptide linker regions. The exons modules of the present combinatorial method can comprise nucleic acid sequences corresponding to naturally occurring exon sequences or naturally occurring exon sequences which have been mutated (e.g. point mutations, truncations, fusions), as well as nucleic acid sequences from "synthetic exons" including sequences of purely random construction, that is, nucleic acid sequences not substantially similar to naturally occurring exon sequences. In some instances, the exon module can correspond to a functional domain, and the module may comprise a number of naturally occurring exon sequences spliced together, with the intron sequences flanking only the exon sequences disposed at the extremity of the module.

However, the term "exon", as used in the present invention, is not limited to proteinencoding sequences, and may comprises nucleic acid sequences of other function, including nucleic acids of "intronic origin" which give rise to, for example, ribozymes or other nucleic acid structure having some defined chemical function. As illustrated below, group II intron domains (e.g. domains I-VI) and group I intron domains (e.g. paired regions P1-P10) can themselves be utilized as "exons", each having flanking intronic sequences that can mediate combinatorial splicing between different group I or group II domains to produce novel catalytic intron structures. In another illustrative embodiment, the exon can comprise a cloning or expression vector into which other nucleic acids are ligated by an intron-mediated trans-splicing reaction.

With respect to generating the protein-encoding exon constructs of the present invention, coding sequences can be isolated from either cDNA or genomic sources. In the instance of cDNA-derived sequences, the addition of flanking intronic fragments to particular portions of the transcript can be carried out to devise combinatorial units having exonic sequences that correspond closely to the actual exon boundaries in the pre-mRNA. Alternatively, the choice of coding sequences from the cDNA clone can be carried out to create combinatorial units having

"exon" portions chosen by some other criteria. For example, as described below with regard to the construction of combinatorial units from either antibody or plasminogen activator cDNA sequences, the criteria for selecting the exon portions of each splicing construct can be based on domain structure or function of a particular portion of the protein.

Several strategies exist for identifying coding sequences in mammalian genomic DNA which can subsequently be used to generate the present combinatorial units. For example, one strategy frequently used involves the screening of short genomic DNA segments for sequences that are evolutionarily conserved, such as the 5' splice site and branch acceptor site consensus sequences (Monaco et al. (1986) Nature 323:646-650; Rommens et al. (1989) Science 245:1059-1065; and Call et al. (1990) Cell 60: 509-520). Alternative strategies involve sequencing and analyzing large segments of genomic DNA for the presence of open reading frames (Fearson et al. (1990) Science 247:49-50), and cloning hypo-methylated CpG islands indicative of 5' transcriptional promoter sequences (Bird et al. (1986) Nature 321:209-213). Yet another technique comprises the cloning of isolated genomic fragments into an intron which is in turn disposed between two known exons. The genomic fragments are identified by virtue of the ability of the inserted genomic sequences to direct alternate splicing which results in the insertion into a mature transcript of at least one genomic-derived exon between the two know exons (Buckler et al. (1991) PNAS 88:4005-4009).

Exons identified from genomic DNA can be utilized directly as combinatorial units by isolating the identified exon and appropriate fragments of the flanking intron sequences normally associated with it. Alternatively, as with the cDNA derived exons, the genomicderived exon can be manipulated by standard cloning techniques (Molecular Biology. A Laboratory Manual, eds. Sambrook, Fritsch and Maniatis (New York: CSH Press, 1989); and Current Protocols in Molecular Biology, Eds. Ausebel et al. (New York: John Wiley & Sons, 1989)) into vectors in which appropriate flanking intronic sequences are added to the exon upon transcription. In yet another embodiment, the reversal of splicing reactions, described below for the various intron groups, can be used to specifically add flanking intron fragments to one or both ends of the exonic sequences, and thereby generate the combinatorial units of the present invention.

Furthermore, generating the splicing units useful in the present combinatorial methods, one skilled in the art will recognize that in the instance of protein-encoding exons,particular attention should be payed to the phase of the intronic fragments. Introns that interrupt the reading frame between codons are known as "Phase 0" introns; those which interrupt the codons between the first and second nucleotides are known as "Phase 1" introns; and those interrupting the codons between the second and third nucleotides are known as "Phase 2" introns. In order to prevent a shift in reading frame upon ligation of two exons, the phase at both the 5' splice site and 3' splice site must be the same. The phase of the flanking intronic fragments can be easily controlled during manipulation, especially when reverse splicing is utilized to add the intronic fragments, as the each insertion site is known. However, as described below, when the variegated population of combinatorial units comprises flanking intronic fragments of mixed phase, particular nucleotides in the intronic sequences can be changed in such a manner as to lower the accuracy of splice site choice. In addition, the splicing reaction conditions can also be manipulated to lower the accuracy of splice site choice.

I. Intronic Sequences

The present invention makes use of the ability of introns to mediate ligation of exons to one and other in order to generate a combinatorial library of genes from a set of discontinuous exonic sequences. This method is not limited to any particular intron or class of introns. By way of example, the intronic sequences utilized can be selected from group I, group II, group III or nuclear pre-mRNA introns. Furthermore, in light of advancements made in delineating the critical and dispensable elements in each of the classes of introns, the present invention can be practiced with portions of introns which represent as little as the minimal set of intronic sequences necessary to drive exon ligation.

Group I introns, as exemplified by the Tetrahymena ribosomal RNA (rRNA) intron, splice via two successive phosphate transfer, transesterification reactions. As illustrated in FIG. 3, the first transesterification is initiated by nucleophilic attack at the 5' junction by the 3' OH of a free guanosine nucleotide, which adds to the 5' end of the intron and liberates the 5' exon with a 3' OH. The second transesterification reaction is initiated by nucleophilic attack at the 3' splice junction by the 3' OH of the 5' exon, which results in exon ligation and liberates the intron.

Group II introns also splice by way of two successive phosphate transfer, transesterification reactions (see FIG. 1). There is, however, one prominent difference between the reaction mechanisms proposed for group I and group II introns. While cleavage at the 5' junction in group I splicing is due to nucleophilic attack by a free guanosine nucleotide, cleavage at the 5' junction in group II splicing is typically due to nucleophilic attack by a 2' OH from within the intron. This creates a lariat intermediate with the 5' end of the intron attached through a 2', 5'-phosphodiester bond to a residue near the 3' end of the intron. Subsequent cleavage at the 3' junction results in exon ligation and liberates the "free" intron in the form of a lariat. The nature of the initiating nucleophile notwithstanding, the two self-splicing mechanism appear quite similar as both undergo 5' junction cleavage first, and subsequently 3' junction cleavage and exon ligation as a consequence of nucleophilic attack by the 5' exon. Furthermore, nuclear pre-mRNA, in similar fashion to group II-intron splicing, also proceed through a lariat intermediate in a two-step reaction.

All three intron groups share the feature that functionally active introns able to mediate splicing can be reconstituted from intron fragments by non-covalent interactions between the fragments (and in some instances other trans-acting factors). Such "trans-splicing" by fragmented introns, as described herein, can be utilized to ligate discontinuous exon sequences to one and other and create novel combinatorial genes. Moreover, autocatalytic RNA (i.e. group I and group II introns) are not only useful in the self-splicing reactions used generate combinatorial libraries, but can also catalyze reactions on exogenous RNA.

The following description of each of the group I, group II, and nuclear pre-mRNA intronic sequences is intended to illustrate the variation that exists in each group of introns. Moreover, the descriptions provide further insight to one skilled in the art to devise exon constructs useful in the present splicing methods, using as little as a minimal set of intronic fragments.

A. Group II Introns

Group II introns, which are classed together on the basis of a conserved secondary structure, are found in organellar genes of lower eukayotes and plants. Like introns in nuclear pre-mRNA, group II introns are excised by a two-step splicing reaction to generate branched circular RNAs, the so-called intron-lariats. A remarkable feature of group II introns is their self-splicing activity in vitro. In the absence of protein or nucleotide cofactors, the intronic RNA catalyzes two successive transesterfication reactions which lead to autocatalytic excision of the intron-lariat from the pre-mRNA and concomitantly to exon ligation. (See FIG. 1).

More than 100 group II intron sequences from fungal and plant mitochondria and plant chloroplasts have been analyzed for conservation of primary sequence, secondary structure and three-dimensional base pairings. Group II introns show considerable sequence homology at their 3' ends (an AY sequence), and have a common G.sub.1 W.sub.2 G.sub.3 Y.sub.4 G.sub.5 motif at their 5' ends, but do not show any other apparent conserved sequences in their interior parts. However, group II introns are generally capable of folding into a distinctive and complex secondary structure typically portrayed as six helical segments or domains (designated herein as domains I-VI) extending from a central hub (see FIG. 2). This core structure is believed to create a reactive center that promotes the transesterification reactions.

However, mutational analysis and phylogenetic comparison indicate that certain elements of the group II intron are dispensable to self-splicing. For example, several group II introns from plants have undergone some rather extensive pruning of peripheral and variable stem structures. Moreover, while the group II intron can be used to join two exons via cis-splicing, a discontinuous group II intron form of trans-splicing can be used which involves the joining of independently transcribed coding sequences through interactions between intronic RNA pieces. In vitro studies have shown that breaks, for example within the loop region of domain IV, can be introduced without disrupting self-splicing. The ability of group II intron domains to reassociate specifically in vivo is evidenced by trans-spliced group II introns, which have been found, for example, in the rps-12 gene of higher plant ctDNA, the psaA gene in Chlamydomonas reinhardtii ctDNA, and the nadl and nad5 genes in higher plant mtDNA (Michel et al. (1989) Gene 82:5-30; and Sharp et al. (1991) Science 254:663). These genes consist of widely separated exons flanked by 5'- or 3'-segments of group II introns split in either domains III or IV. The exons at different loci are transcribed into separate precursor RNAs, which are trans-spliced, presumably after the association of the two segments of the group II intron. Moreover, genetic analysis of trans-splicing of the Chlamydomonas reinhardtii psaA gene has demonstrated that the first intron of this gene is split into three segments. The 5' exon is flanked by parts of domain I and the 3' exon by parts of domains IV to VI, respectively. The middle segment of the intron is encoded at a remote locus, tscA, and consists of the remainder of domains I to IV. This tscA segment can apparently associate with the other two intron segments to reconstitute an intron capable of splicing the two exons (Goldschmidt et al. (1991) Cell 65:135-143).

The functional significance to self-splicing of certain control structural elements have been further deduced by analysis of minimal trans-splicing sets, and found to generally comprise an exon-binding site and intron-binding site, a structural domain V, and (though to lesser extent) a "branch-site" nucleotide involved in lariat formation. Domain I contains the exon-binding sequences. Domain VI is a helix containing the branch site, usually a bulged A residue. Domain

V, the most highly conserved substructure, is required for catalytic activity and binds to at least a portion of domain I to form the catalytic core.

The 5' splice sites of group II introns are defined by at least three separate tertiary base pairing contacts between nucleotides flanking the 5' splice site and nucleotides in substructures of domain I. The first interaction involves a loop sequence in the D sub-domain of domain I (exon binding site 1 or EBS 1) that base pairs with the extreme 3' end of the 5' exon (intron binding site 1 or IBS 1). The second interaction involves the conserved dinucleotide --G.sub.3 Y.sub.4 --(designated .epsilon.) that base pairs with a dinucleotide in the C1 subdomain of domain I (designated .epsilon.'). The third interaction involves base pairing between intron binding site 2 (IBS 2), a sequence located on the 5' side of IBS 1, with exon binding site 2 (EBS 2), a loop sequence of the D subdomain of domain I near EBS 1. Of the two exon-binding sites identified in group II introns, only EBS 1 is common to all group II members. The EBS 1 element comprises a stretch of 3 to 8 consecutive residues, preferably 6, located within domain I, which are complementary to the last 3 to 8 nucleotides of the 3' exon end of the 5' exon. The EBS 2-IBS 2 pairing also typically consists of two 4-8 nucleotide stretches. Its exonic componebt (IBS 2) lies from 0 to 3 nucleotides upstream from the IBS 1 element, and the intronic component (EBS 2) also lies within domain I. However, while IBS 2-EBS 2 pairing can improve the efficiency of 5' splice site use, particularly in trans-, it is subject to many more variations from the IBS 1-EBS 1 interaction, such as reduced length, presence of bulging nucleotides or a mismatch pair. Disrupting the IBS 2-EBS 2 pairing, in the Sc.a5 group II intron for example, is essentially without effect on the normal splicing reaction, and in at least twelve group II introns analyzed, the IBS 2-EBS 2 interaction seems to be missing altogether and is apparently less important than the IBS 1-EBS 2 interaction. As already noted, only that pairing is absolutely constant in (typical) group II introns, and always potentially formed at cryptic 5' splice sites.

Further studies, while confirming that the EBS 1-IBS 1 base pairing is necessary for activation of the 5' junction, indicate that this interaction alone is not always sufficient for unequivocal definition of the cleavage site. It has been established that altering the first nucleotide of the group II intron (e.g., G.sub.1 of G.sub.1 W.sub.2 G.sub.3 Y.sub.4 G.sub.5) can reduce the self-splicing rate in vitro. Characterization of the products of self-splicing from G.sub.1 .fwdarw.N mutant transcripts have demonstrated that the relative order of function is G>U>A>C. It is also suggested that the 5' G of the intron helps to position the cleavage site precisely (Wallasch et al. (1991) Nuc. Acid Res. 19:3307-3314). For example, the presence of an additional adenosine following IBS 1can lead to ambiguous hydrolytic cleavages at the 5' intron/exon boundary. As described herein, such ambiguity can be used to address exon phasing.

Another well conserved feature of group II introns is the bulging A located 7 to 8 nt upstream from the 3' intron-exon junction on the 3' side of helix VI. This is the nucleotide which participates in the long range, 2'-5' lariat bond (Van der Veen et al. (1986) Cell 44:225-234; Schmelzer and Schweyen (1986) Cell 46:557-565; Jacquier and Michel (1987) Cell 50:17-29; Schmelzer and Muller (1987) Cell 51:753-762). Evidence from electron microscopy, attempts at reverse transcription of circular introns, and treatment with the 2',5'-phosphodiesterase of HeLa cells indicate that group II introns are excised as lariats (Van der Veen et al. (1986) Cell 44:225-234; Schmidt et al. (1987) Curr. Genet. 12:291-295; Koller et al. (1985) Embo J. 4:2445-2450). However, lariat formation is not absolutely essential for correct exon ligation to occur. Cleavage

at the 5' splice site, presumably mediated by free hydroxide ions rather than a 2'--OH group, followed by normal exon ligation, has been observed both in trans-splicing reactions (Jacquier and Rosbash (1986) Science 234:1099-1104; and Koch et al. (1992) Mol. Cell Biol. 12:1950-1958) and, at high ionic strength, in cis-splicing reactions with molecules mutated in domain VI (Van der Veen et al. (1987) Embro J. 12:3827-3821). Also, several group II introns lack a bulging A on the 3' side of helix VI. For instance, all four CP tRNA-VAL introns of known sequence have a fully paired helix VI, and their 7th nucleotide upstream from the 3' intron-exon junction is a G, not an A. Furthermore, correct lariat formation has been observed with a mutant of intron Sc.b1 whose helix VI should be fully paired, due to the insertion of an additional nucleotide (a U) at the site facing the normally bulging A (Schmelzer and Muller (1987) Cell 51:753-762).

Perhaps one of the best conserved structural elements of group ll introns is domain V. The typical domain V structure contains 32-34 nucleotides and is predicted to fold as a hairpin. The hairpin is typically an extended 14 base pair helix, capped by a four base loop involving 15-18, and punctuated by a 2 base bulge at positions 25 and 26. Comparative sequence analysis (Michel et al. (1989) Gene82:5-30) has shown that group II introns can generally be classified into one of two classes (e.g. group IIA and IIB). FIG. 8 shows the consensus sequences of domain V for each of the IIA and IIB introns. Base pairs that are highly conserved are indicated by solid lines. Dashed lines indicate less well conserved base pair interactions. The unpaired loop at the apex of the hairpin is typically an NAAA sequence, where N is most often a G for IIA introns. Nucleotides which are highly conserved are circled, while less conserved nucleotides are uncircled. A black dot indicates a lack of discernible sequence consensus.

Degenerate group II introns can be functional despite lacking some domains. Euglena ctDNA, for example, contains a large number of relatively short group II introns which sometimes lack recognizable cognates of domain II, III, or IV. The view that the only group II structures required for splicing activities are domains I and V is supported by a detailed mutational analysis of a yeast mitochondrial group II intron in which various domains were deleted, either singly or in combinations. (Koch et al. (1992) Mol. Cell. Biol. 12:1950-1958). For example, the removal or disruption of the domain VI helix simply reduces 3' splice site fidelity and reaction efficiency. This analysis has led to the belief that domain V probably interacts with domain I to activate the 5' splice site, since a transcript lacking domains II-IV, and VI, but retaining domain I and domain V was capable of specific hydrolysis of the 5' splice junction.

With regard to 3' splice-site selection, two weak contacts are believed to play a role in defining the 3' splice-site but are, however, not essential to splicing. The first of these contacts is a lone base pair, termed .gamma./.gamma.', between the 3' terminal nucleotide of the introns and a single base between domains II and III. (Jacquier et al. (1990) J. Mol Biol. 13:437-447). A second single base pair interaction, termed the internal guide, has been defined between the first base of the 3' exon and the nucleotide adjacent to the 5' end of EBS 1 (Jacquier et al. (1990) J Mol. Biol. 219:415-428).

In addition to the ability of autocatalytic RNAs such as group I and group II introns to excise themselves from RNA and ligate the remaining exon fragments, ample evidence has accumulated demonstrating that the autocatalytic RNAs can also catalyze their integration into

exogenous RNAs. For example, both group I and group II introns can integrate into foreign RNAs by reversal of the self-splicing reactions. The mechanism of the group II intron reverse-splicing reaction is shown in FIG. 1. In the first step of the reverse reaction, the attack of the 3' OH group of the intron 3' terminus at the junction site of the ligated exons yields a splicing intermediate, the intron-3' exon lariat, and the free 5' exon. In the second step, the 5' exon which is still bound to the lariat via the IBS 1/EBS 1 base pairing can attack the 2'-5' phosphodiester bond of the branch. This transesterification step leads to reconstitution of the original precursor. The analogous reaction of the intron with a foreign RNA harboring an IBS 1 motif results in site-specific integration downstream of the IBS 1 sequence.

The exon constructs of the present invention, whether comprising the group II intronic sequences described above or the group I or nuclear pre-mRNA intronics described below, can be generated as RNA transcripts by synthesis in an in vitro transcription system using well known protocols. For example, RNA can be transcribed from a DNA template containing the exon construct using a T3 or T7 RNA polymerase, in a buffer solution comprising 40 mm Tris-HCl (pH 7.5), 6 mM MgCl.sub.2, 10 mM dithiothreitol, 4 mM spermidine and 500 mM each ribonucleoside triphosphate. In some instances, it will be desirable to omit the spermidine from the transcription cocktail in order to inhibit splicing of the transcribed combinatorial units.

Several reaction conditions for facilitating group II-mediated splicing are known. For example, the reaction can be carried out in "Buffer C" which comprises 40 mM Tris-HCl (pH 7.0), 60 mM MgCl.sub.2, 2 mM spermidine, and 500 mM KCl (Wallasch et al. (1991) Nuc. Acid Res. 19:3307-3314; and Suchy et al. (1991) J. Mol. Biol. 222:179-187); or "Buffer S" which comprises 70 mM Tris-SO.sub.4 (pH 7.5) 60 mM MgSO.sub.4, 2 mM spermidine, and 500 mM (NH.sub.4).sub.2 SO.sub.4 (Morl et al. (1990) Nuc. Acid Res. 18:6545-6551; and Morl et al. (1990) Cell 60:629-636). The group II ligation reactions can be carried out, for instance, at 45.degree. C., and the reaction stopped by EtOH precipitation or by phenol:chloroform (1:1) extraction. Suitable reaction conditions are also disclosed in, for example, Jacquier et al. (1986) Science 234:1099-1104; Franzer et al. (1993) Nuc. Acid Res. 21:627-634; Schmelzer et al. (1986) Cell 46:557-565; Peebles et al. (1993) J. Biol. Chem. 268:11929-11938; Jarrell et al. (1988) J Biol. Chem. 263:3432-3439; and Jarrell et al. (1982) Mol. Cell Biol. 8:2361-2366. Moreover, manipulation of the reaction conditions can be used to favor certain reaction pathways, such as a reverse-splicing reaction (e.g., by increasing the MgSO.sub.4, concentration to 240 mM in Buffer S); bypassing the need for a branch nucleotide acceptor (e.g. high salt); and decreasing the accuracy of splice-site choice (Peebles et al. (1987) CSH Symp. Quant. Biol. 52:223-232).

B. Group I Introns

Group I introns are present in rRNA, tRNA, and protein-coding genes. They are particularly abundant in fungal and plant mitochondrial DNAs (mtDNAs), but have also been found in nuclear rRNA genes of Tetrahymena and other lower eukaryotes, in chloroplast DNAs (ctDNAs), in bacteriophage, and recently in several tRNA genes in eubacteria.

As first shown for the Tetrahymena large rRNA intron, group I introns splice by a mechanism involving two transesterification reactions initiated by nucleophilic attack of guanosine at the 5'

splice site (See FIG. 3). The remarkable finding for the Tetrahymena intron was that splicing requires only guanosine and Mg.sup.2+. Because bond formation and cleavage are coupled, splicing requires no external energy source and is completely reversible. After excision, some group I introns circularize via an additional transesterfication, which may contribute to shifting the equilibrium in favor of spliced products.

The ability of group I introns to catalyze their own splicing is related to their highly conserved secondary and tcrtiary structures. The folding of the intron results in the formation of an active site juxtaposing key rcsidues that are widely separated in primary sequence. This RNA structure catalyzes splicing by bring the 5' and 3' splice sites and guanosine into proximity and by activating the phosphodiester bonds at the splice sites. Different group I introns have relatively little sequence similarity, but all share a series of the short, conserved sequence elements P, Q, R, and S. These sequence elements always occur in the same order and basepair with one another in the folded structure of the intron (see FIG. 4). Element R [consensus sequence (C/G)YUCA(GA/AC)GACUANANG (SEQ ID NO. 4)] and S [consensus AAGAUA-GUCY (SEQ ID No: 5)] are the most highly conserved sequences within group I introns, and typically serve as convenient "landmarks" for the identification of group I introns. The boundaries of group I introns are marked simply by a U residue at the 3' end of the 5' exon and a G residue at the 3' end of the intron. (see, for example, Michel et al. (1990) J Mol Biol 216:585-610; Cech, TR (1990) Annu Rev Biochem 59:543-568; Cech, TR (1988) Gene 73:259-271; Burke (1989) Methods in Enzymology 190:533-545; and Burke et al. (1988) Gene 73:273-294)

The conserved group I intron secondary structure was deduced from phylogenetic comparisons, and specific features have been confirmed by analysis of in vivo and in vitro mutations and by structure mapping. The structure, shown in FIG. 4, consists of a series of paired regions, denoted P1-P10, separated by single-stranded regions (denoted J) or capped by loops (denoted L), from the core of the structure. The fundamental correctness of the model is supported by the observation that a vast number of group I intron sequences can be folded into this basic structure.

P1and P10, which contain the 5' and 3' splice sites, respectively, are formed by base pairing between an internal guide sequence (IGS), generally located just downstream of the 5' splice site, and exon sequences flanking the splice sites. Group I introns have been classified into four major subgroups, designated IA to ID, based on distinctive structural and sequence features. Group IA introns, for example, contain two extra pairings, P7.1/P7.1a or P7.1/P7.2, between P3 and P7, whereas many group IB and IC introns may contain additional sequences, including open reading frames (ORFs), in positions that do not disrupt the conserved core structure. Indeed, many of the peripheral stem-loops can be completely deleted without major loss of splicing function. For example, the phage T4 sunY intron has been re-engineered to contain as few as 184 nucleotides while still retaining greater than 10-percent activity. Presumably, if the criterion for activity were lowered, the minimal size one could achieve would be decreased.

The region of the Tetrahymena intron required for enzymatic activity, the catalytic core, consists of P3, P4, P6, P7, P8, and P9.0. Mutation of a nucleotide involved in one of these core structural elements typically decreases the maximum velocity of splicing, increase K.sub.m for guanosine, or both. In those instances where the primary importance of the nucleotide is its contribution to the formation of a duplex region, a second-site mutation that restores basepairing also restores

splicing function. Studies using Fe(II)-EDTA, a reagent that cleaves the sugar-phosphate backbone, have shown that parts of the core are buried in the structure inaccessible to the solvent, that Mg.sup.2+ is necessary for folding of the intron, and that individual RNA domains fold in a specific order as Mg.sup.2+ is increased. All group I introns have fundamentally similar core structures, but subgroup-specific structures such as P7.1, P7.2, and P5abc appear to participate in additional interactions that stabilize the core structure in different ways (Michel et al. (1990) J Mol Biol 216:585-610; and Michel et al. (1992) Genes & Dev 6:1373-1385).

A three dimensional model of the group I intron catalytic core has been developed by Michel and Westhof (Michel et al. (1990) J Mol Biol 216:585-610) through comparative sequence analysis. In the Michel-Westhof model, the relative orientation of the two helices is constrained by a previously proposed triple helix involving parts of J3/4-P4-P6-J6/7 and by potential tertiary interactions identified by co-variation of nucleotides that are not accounted for by secondary structure. A number of these binding sites accounts for the known splicing mechanism, which requires appropriate alignments of guanosine and the 5' and 3' exons in the first and second steps of splicing. Deoxynucleotide and phosphorothioate substitution experiments suggest that functionally important Mg.sup.2+ ions are coordinated at specific positions around the active site (e.g., P1 and J8/7) where they may function directly in phosphodiester bond cleavage (Michel et al. (1990) J Mol Biol 216:585-610; and Yarus, M (1993) FASEB J 7:31-9). Basic features of the predicted three-dimensional structure have been supported by mutant analysis in vitro and by the use of specifically positioned photochemical cross-linking and affinity cleavage reagents.

The 5' and 3' splice sites of group I introns are substrates that are acted on by the catalytic core, and they can be recognized and cleaved by the core when added on separate RNA molecules (Cech, TR (1990) Annu Rev Biochem 59:543-568). In group I introns the last 3-7 nucleotides of the 5' exon are paired to a sequence within the intron to form the short duplex region designated P1. The intron-internal portion of P1 is also known as the 5' exon-binding site and as a portion of the internal guide sequence, IGS. The P1s of different group I introns vary widely in sequence. Neither the sequence nor length of P1 is fixed, but the conserved U at the 3' end of the 5' exon always forms a wobble base pair with a G residue in the IGS (FIG. 4). The conserved U:G is one important recognition element that defines the exact site of guanosine attack. In general, other base combinations do not substitute well. One exception is C:G, which maintains the accuracy of splicing but decreases the Kcat/Km by a factor of 100. Another exception is C:A; the ability of this pair to substitute well for U:G has been interpreted as an indication that disruption of P1 by a wooble base pair is a key element in recognition of the splice site. Position within the P1 helix is another determinant of 5' splice site. Analysis of in vitro mutants has shown that the distance of the U:G pair from the bottom of the P1 helix is critical for efficient cleavage in the Tetrahymena intron and that J1/2 and P2 also play a role in the positioning of P1 relative to the core (Michel et al. (1990) J Mol Biol 216:585-610; Young et al. (1991) Cell 67:1007-1019; and Salvo et al. (1992) J Biol Chem267:2845-2848). The U:G pair is most efficiently used when located 4-7 base pairs from the base of the P1.

The positioning of the 3' splice site in group I introns depends on at least three interactions, whose relative importance differs in different introns. These are the P10 pairing between the IGS and the 3' exon, binding of the conserved G residue at the 3' end of the intron to the G-binding site in the second step of splicing, and an additional interaction, P9.0, which involves base paring

between the two nucleotides preceding the terminal G of the intron and two nucleotides in J7/9 (Cech, TR (1990) Annu Rev Biochem 59:543-568).

Group I introns have Km values for guanosine that are as low as 1 .mu.M and readily discriminate between guanosine and other nucleosides. The major component of the guanosine-binding site corresponds to a universally conserved CG pair in P7. Guanosine was initially proposed to interact with this base pair via formation of a base triple, but the contribution of neighboring nucleotides and the binding of analogs are also consistent with a model in which guanosine binds axially to the conserved G and flanking nucleotides. The guanosine-binding site of group I introns can also be occupied by the guanidino groups of arginine or antibiotics, such as streptomycin, which act as competitive inhibitors of splicing (von Ahsen et al. (1991) Nuc Acids Res 19:2261-2265).

Group I introns can also be utilized in both trans-splicing and reverse-splicing reactions. For example, the ribozyme core of a group I intron can be split in L6, and through intermolecular complementation, a functional catalytic core can be reassembled from intronic fragments (i.e. P1-6.5 and P6.5-10) on separately transcribed molecules (Galloway et al. (1990) J. Mol. Biol. 211:537-549).

Furthermore, as described for group II intron constructs, combinatorial units comprising group I introns can be transcribed from DNA templates by standard protocols. The group I self-splicing reaction has an obligatory divalent cation requirement, which is commonly met by Mg.sup.2+ . The reaction can in fact be stopped using a chelating agent such as EDTA. The group I-mediated splicing of exonic sequences can be carried out, for example, in a buffer comprising 100 mM (NH.sub.4).sub.2 SO.sub.4, 50 mM HEPES (pH 7.5), 10 mM MgCl.sub.2, and 25 .mu.M GTP, at a temperature of 42.degree. C. (Woodson et al. (1989) Cell 57:335-345). In another embodiment, the reaction buffer comprises 50 mM Tris-HCl (pH 7.5), 50 mM NH.sub.4 Cl, 3 mM MgCl.sub.2, 1 mM spermidine, and 100 mM GTP, and the reaction proceeds at 55.degree. C. (Salvo et al. (1990) J. Mol. Biol. 211:537-549). To form the reverse-splicing reaction, the Mg.sup.2+ concentration can be increased (e.g., to 25 mM) and the GTP omitted. Typically, the reversal of splicing reaction is favored by high RNA concentrations, high magnesium and temperature, and the absence of guanosine. Other examples of useful reaction conditions for group I intron splicing can be found, for example, in Mohr et al. (1991) Nature 354:164-167; Guo et al. (1991) J Biol. Chem. 266:1809-1819; Kittle et al. (1991) Genes Dev. 5:1009-1021; Doudna et al. (1989) PNAS 86:7402-7406; and Pattanju et al. (1992) Nuc. Acid Res. 20:5357-5364.

The efficiency of splicing of group II and group I introns can often be improved by, and in some instances may require, the addition of protein and/or RNA co-factors, such as maturases. (Michel et al. (1990) J Mol.Biol. 216:585-610; Burke et al. (1988) Gene 71:259-271; and Lambowitz et al. (1990) TIBS 15:440-444). This can be especially true when more truncated versions of these introns are used to drive ligation by trans-splicing, with the maturase or other co-factor compensating for structural defects in the intron structure formed by intermolecular complementation by the flanking intron fragments. Genetic analysis of mitochondrial RNA splicing in Neurospora and yeast has shown, for example, that some proteins involved in splicing of group I and group II introns are encoded by host chromosomal genes, whereas others are

encoded by the introns themselves. Several group I and group II introns in yeast mtDNA, for instance, encode maturases that function in splicing the intron that encodes them. These include group I introns Cob-I2, -I3, and I4, and group II introns cox1-I1 and -I2. Thus, the conditions for splicing of group I and group II introns can further comprise maturases and other co-factors as necessary to form a functional intron by the flanking intron sequences.

C. Nuclear pre-mRNA introns

Nuclear pre-mRNA splicing, like group II intron-mediated splicing, also proceeds through a lariat intermediate in a two-step reaction. In contrast to the highly conserved structural elements that reside within group II introns, however, the only conserved features of nuclear pre-mRNA introns are restricted to short regions at or near the splice junctions. For instance, in yeast these motifs are (i) a conserved hexanucleotide at the 5' splice, (ii) an invariant heptanucleotide, the UACUAAC box, surrounding the branch point A (underlined), and (iii) a generally conserved enrichment for pyrimidine residues adjacent to an invariant AG dinucleotide at the 3' splice site.

Two other characteristics of nuclear pre-mRNA splicing in vitro that distinguish it from autocatalytic splicing are the dependence on added cell-free extracts and the requirement for adenosine triphosphate (ATP). Once in vitro systems had been established for mammalian and yeast pre-mRNA splicing, it was found that a group of trans-acting factors, predominately made up of small nuclear ribonucleoprotein particles (snRNP's) containing U1, U2, U4, U5 and U6 RNA's was essential to the splicing process. Together with the discovery of autocatalytic introns, the demonstration that snRNAs were essential, trans-acting components of the spliceosome argued strongly that group II self-splicing and nuclear pre-mRNA splicing occurred by fundamentally equivalent mechanisms. According to this view, the snRNAs compensate for the low information content of nuclear introns and, by the formation of intermolecular RNA-RNA interactions, achieve the catalytic capability inherent in the intramolecular structure of autocatalytic introns.

As illustrated in FIG. 9A, consensus sequences of the 5' splice site and at the branchpoint are recognized by base pairing with the U1 and U2 snRNP's, respectively. The original proposal that the U1 RNA interacted with the 5' splice site was based solely on the observed nine-base-pair complementarity between the two mammalian sequences (Rogers et al. (1980) Nature 283:220). This model has since been extensively verified experimentally (reviewed in Steitz et al., in Structure and Function of Major and Minor snRNP Particles, M. L. Bimstiel, Ed. (Springer-Verlag, N.Y., 1988)). Demonstration of the Watson-Crick interactions between these RNAs was provided by the construction of compensatory base pair changes in mammalian cells (Zhuang et al. (1986) Cell 46:827). Subsequently, suppressor mutations were used to prove the interaction between U1 and the 5' splice site in yeast (Seraphin et al. (1988) EMBO J 7:2533).

The base pairing interaction between U2 and sequences surrounding the branchpoint was first tested in yeast (Parker et al. (1987) Cell 49:229), where the strict conservation of the branchpoint sequence readily revealed the potential for complementarity. The branchpoint nucleotide, which carries out nucleophilic attack on the 5' splice site, is thought to be unpaired (FIG. 9A), and is analogous to the residue that bulges out of an intramolecular helix in domain VI of group II introns. The base pairing interaction between U2 and the intron has also been demonstrated

genetically in mammalian systems (Zhaung et al. (1989) Genes Dev. 3:1545). In fact, although mammalian branchpoint sequences are notable for their deviation from a strict consensus, it has been demonstrated that a sequence identical to the invariant core of the yeast consensus, CUAAC is the most preferred (Reed et al. (1989) PNAS 86:2752).

Genetic evidence in yeast suggests that the intron base pairing region at the 5' end of U1 RNA per se is not sufficient to specify the site of 5' cleavage. Mutation of the invariant G at position 5 of the 5' splice site not only depresses cleavage efficiency at the normal GU site but activates cleavage nearby; the precise location of the aberrant site varies depending on the surrounding context (Jacquier et al. (1985) Cell 43:423; Parker et al. (1985) Cell 41:107; and Fouser et al. (1986) Cell 45:81). Introduction of a U1 RNA, the sequence of which has been changed to restore base pairing capability at position 5, does not depress the abnormal cleavage event; it enhances the cleavage at both wild-type and aberrant sites. These results indicate that the complementarity between U1 and the intron is important for recognition of the splice-site region but does not determine the specific site of bond cleavage (Seraphin et al. (1988) Genes Dev. 2:125; and Seraphin et al. (1990) Cell 63:619).

With regard to snRNPs, genetic experiments in yeast have revealed that the U5 snRNP is an excellent candidate for a trans-acting factor that functions in collaboration with U1 to bring the splice sites together in the spliceosome. U5 is involved in the fidelity of the first and the second cleavage-ligation reactions. For example, a number of U5 mutants exhibit a distinct spectrum of 5' splice-site usage; point mutations with the invariant nine-nucleotide loop sequence (GCCUUUUAC) in U5 RNA allows use of novel 5' splice sites when the normal 5' splice site was mutated. For instance, splicing of defective introns was restored when positions 5 or 6 of the invariant U5 loop were mutated so that they were complementary to the nucleotides at positions 2 and 3 upstream of the novel 5' splice site. Likewise, mutational analysis has demonstrated the role of the U5 loop sequence in 3' splice site activation. For example, transcripts which are defective in splicing due to nucleotide changes in either one of the first two nucleotides of the 3' exon were subsequently rendered functional by mutations in positions 3 or 4 of the U5 loop sequence which permitted pairing with the mutant 3' exon. (See Newman et al. (1992) Cell 68:1); and Newman et al. (1991) Cell 65:115). It is suggested that first U1 base pairs with intron nucleotides at the 5' splice site during assembly of an early complex (also including U2). This complex is joined by a tri-snRNP complex comprising U4, U5 and U6 to form a Holliday-like structure which serves to juxtaposition the 5' and 3' splice sites, wherein U1 base pairs with intronic sequences at both splice site. (Steitz et al. (1992) Science 257:888-889).

While each of the U1, U2 and U5 snRNPs appear to be able to recognize consensus signals within the intron, no specific binding sites for the U4-U6 snRNP has been identified. U4 and U6 are well conserved in length between yeast and mammals and are found base paired to one another in a single snRNP (Siliciano et al. (1987) Cell 50:585). The interaction between U4 and U6 is markedly destablized specifically at a late stage in spliceosome assembly, before the first nucleolytic step of the reaction (Pikienly et al. (1986) Nature 324: 341; and Cheng et al. (1987) Genes Dev. 1:1014). This temporal correlation, together with an unusual size and sequence conservation of U6, has lead to the understanding that the unwinding of U4 from U6 activates U6 for participation in catalysis. In this view, U4 would function as an antisense negative regulator, sequestering U6 in an inert conformation until it is appropriate to act (Guthrie et al.

(1988) Annu Rev. Genet. 22:387). Recent mutational studies demonstrate a functional role for U6 residues in the U4-U6 interaction domain in addition to base pairing (Vanken et al. (1990) EMBO J 9:3397; and Madhani et al. (1990) Genes Dev. 4:2264).

Mutational analysis of the spliceosomal RNAs has revealed a tolerance of substitutions or, in some cases, deletion, even of phylogenetically conserved residues (Shuster et al. (1988) Cell 55:41; Pan et al. (1989) Genes Dev. 3:1887; Liao et al. (1990) Genes Dev. 4:1766; and Jones et al. (1990) EMBO J 9:2555). For example, extensive mutagenesis of yeast U6 has been carried out, including assaying the function of a mutated RNA with an in vitro reconstitution system (Fabrizio et al. (1990) Science 250:404), and transforming a mutagenized U6 gene into yeast and identifying mutants by their in vivo phenotype (Madhani et al. (1990) Genes Dev. 4:2264). Whereas most mutations in U6 have little or no functional consequence (even when conserved residues were altered), two regions that are particularly sensitive to nucleotide changes were identified: a short sequence in stem I (CAGC) that is interrupted by the S. pombe intron, and a second, six-nucleotide region (ACAGAG) upstream of stem I.

As described above for both group I and group II introns, exonic sequences derived from separate RNA transcripts can be joined in a trans-splicing process utilizing nuclear pre-mRNA intron fragments (Konarska et al. (1985) Cell 42:165-171; and Solnick (1985) Cell 42:15-164). In the trans-splicing reactions, an RNA molecule, comprising an exon and a 3' flanking intron sequences which includes a 5' splice site, is mixed with an RNA molecule comprising an exon and 5' flanking intronic sequences, including a 3' splice site, and a branch acceptor site. As illustrated in FIGS. 9B and 9C, upon incubation of the two types of transcripts (e.g. in a cell-free splicing system), the exonic sequences can be accurately ligated. In a preferred embodiment the two transcripts contain complementary sequences which allow basepairing of the discontinuous intron fragments. Such a construct, as FIG. 9B depicts, can result in a greater splicing efficiency relative to the scheme shown in FIG. 9C in which no complementary sequences are provided to potentiate complementation of the discontinuous intron fragments.

The exon ligation reaction mediated by nuclear pre-mRNA intronic sequences can be carried out in a cell-free splicing system. For example, combinatorial exon constructs can be mixed in a buffer comprising 25 mM creatine phosphate, 1 mM ATP, 10 mM MgCl.sub.2, and a nuclear extract containing appropriate factors to facilitate ligation of the exons (Konarska et al. (1985) Nature 313:552-557; Krainer et al. (1984) Cell 36:993-1005; and Dignam et al. (1983) Nuc. Acid Res. 11:1475-1489). The nuclear extract can be substituted with partially purified spliceosomes capable of carrying out the two transesterfication reactions in the presence of complementing extracts. Such spliceosomal complexes have been obtained by gradiant sedimentation (Grabowski et al. (1985) Cell 42:345-353; and Iin et al. (1987) Genes Dev. 1:7-18), gel filtration chromatography (Abmayr et al. (1988) PNAS 85:7216-7220; and Reed et al. (1988) Cell 53:949-961), and polyvinyl alcohol precipitation (Parent et al. (1989) J. Mol. Biol. 209:379-392). In one embodiment, the spliceosomes are activated for removal of nuclear pre-mRNA introns by the addition of two purified yeast "pre-mRNA processing" proteins, PRP2 and PRP16 (Kim et al. (1993) PNAS 90:888-892; Yean et al. (1991) Mol. Cell Biol. 11:5571-5577; and Schwer et al. (1991) Nature 349:494-499).

II. Trans-splicing Combination of Exons

In one embodiment of the present combinatorial method, the intronic sequences which flank each of the exon modules are chosen such that gene assembly occurs in vitro through ligation of the exons, mediated by a trans-splicing mechanism. Conceptually, processing of the exons resembles that of a fragmented cis-splicing reaction, though a distinguishing feature of trans-splicing versus cis-splicing is that substrates of the reaction are unlinked. As described above, breaks in the intron sequence can be introduced without abrogating splicing, indicating that coordinated interactions between different portions of a functional intron need not depend on a covalent linkage between those portions to reconstitute a functionally-active splicing structure. Rather, the joining of independently transcribed coding sequences results from interactions between fragmented intronic RNA pieces, with each of the separate precursors contributing to a functional trans-splicing core structure.

The present trans-splicing system provides an active set of transcripts for trans-splicing wherein the flanking intronic sequences can interact to form a reactive complex which promotes the transesterification reactions necessary to cause the ligation of discontinuous exons. In one embodiment, the exons are flanked by portions of one of a group I or group II intron, such that the interaction of the flanking intronic sequences is sufficient to produce an autocatalytic core capable of driving ligation of the exons in the absence of any other factors. While the accuracy and/or efficiency of these autocatalytic reactions can be improved, in some instances, by the addition of trans-acting protein or RNA factors, such additions are not necessary.

In another embodiment, the exon modules are flanked by intronic sequences which are unable, in and of themselves, to form functional splicing complexes without involvement of at least one trans-acting factor. For example, the additional trans-acting factor may compensate for structural defects of a complex formed solely by the flanking introns. As described above, domain V of the group II intron class can be removed from the flanking intronic sequences, and added instead as a trans-acting RNA element. Similarly, when nuclear pre-mRNA intron fragments are utilized to generate the flanking sequences, the ligation of the exons requires the addition of snRNPs to form a productive splicing complex.

In an illustrative embodiment, the present combinatorial approach can make use of group II intronic sequences to mediate trans-splicing of exons. For example, as depicted in FIG. 5, internal exons can be generated which include domains V and VI at their 5' end, and domains I-III at their 3' end. The nomenclature of such a construct is (IVS5,6)Exon(IVS1-3), representing the intron fragments and their orientation with respect to the exon. Terminal exons are likewise constructed to be able to participate in trans-splicing, but at only one end of the exon. A 5' terminal exon, in the illustrated group II system, is one which is flanked by domains I-III at its 3' end [Exon(IVS1-3)] and is therefore limited to addition of further exonic sequences only at that end; and a 3' terminal exon is flanked by intron sequences (domains V and VI) at only its 5' end [(IVS5,6)Exon]. Under conditions which favor transsplicing, the flanking intron sequences at the 5' end of one exon and the 3' end of another exon will associate to form a functionally active complex by intermolecular complementation and ligate the two exons together.

In another embodiment of the present trans-splicing combinatorial method, the exons, as initially admixed, lack flanking intronic sequences at one or both ends, relying instead on a subsequent

addition of flanking intronic fragments to the exons by a reverse-splicing reaction. Addition of the flanking intron sequences, which have been supplemented in the exon mixture, consequently activates an exon for trans-splicing. FIG. 6 illustrates how the reverse-splicing reaction of group II introns can be used to add domains I-IV to the 3' end of an exon as well as domains I-III to the 5' end of an exon. As shown in FIG. 6, the reversal reaction for branch formation can mediate addition of 3' flanking sequences to an exon. For example, exon modules having 5' intron fragments (e.g. domains V-VI) can be mixed together with little ligation occurring between exons. These exons are then mixed with a 2'-5' Y-branched intron resembling the lariate-IVS, except that the lariat is discontinuous between domain IV and V. The reverse-splicing is initiated by binding of the IBS 1 of the 5' exon to the EBS 1 of the Y-branched intron, followed by nucleophilic attack by the 3'--OH of the exon on the 2'-5' phosphodiester bond of the branch site. This reaction, as depicted in FIG. 6, results in the reconstitution of the 5' splice-site with a flanking intron fragment comprising domains I-III.

While FIG. 6 depicts both a 5' exon and 3' exon, the reverse splicing reaction can be carried out without any 3' exon, the IBS sequence being at the extreme 3' end of the transcript to be activated. Alternatively, to facilitate addition of 5' flanking sequences, an exon can be constructed so as to further include a leader sequence at its 5' end. As shown in FIG. 6, the leader (e.g. the 5' exon) contains an IBS which defines the splice junction between the leader and "mature" exon. The leader sequence can be relatively short, such as on the order of 2-3 amino acid residues (e.g. the length of the IBS). Through a reverse self-splicing reaction using a discontinuous 2'-5' branched intron, the intronic sequences can be integrated at the splice junction by reversal of the two transterfication steps in forward splicing. The resulting product includes the mature exon having a 5' flanking intron fragment comprising domains V and VI.

Addition of intronic fragments by reverse-splicing and the subsequent activation of the exons presents a number of control advantages. For instance, the IBS:EBS interaction can be manipulated such that a variegated population of exons is heterologous with respect to intron binding sequences (e.g. one particular species of exon has a different IBS relative to other exons in the population). Thus, sequential addition of intronic RNA having discrete EBS sequences can reduce the construction of a gene to non-random or only semi-random assembly of the exons by sequentially activating only particular combinatorial units in the mixture. Another advantage derives from being able to store exons as part of a library without self-splicing occurring at any significant rate during storage. Until the exons are activated for trans-splicing by addition of the intronic sequences to one or both ends, the exons can be maintained together in an effectively inert state.

When the interactions of the flanking introns are random, the order and composition of the internal exons of the combinatorial gene library generated is also random. For instance, where the variegated population of exons used to generate the combinatorial genes comprises N different internal exons, random trans-splicing of the internal exons can result in N.sub.y different genes having y internal exons. Where 5 different internal exons are used (N=5) but only constructs having one exon ligated between the terminal exons are considered (i.e. y=1) the present combinatorial approach can produce 5 different genes. However, where y=6, the combinatorial approach can give rise to 15,625 different genes having 6 internal exons, and 19,530 different genes having from 1 to 6 internal exons (e.g. N.sup.1 +N.sup.2 . . . N.sup.-1

+N.sup.y). It will be appreciated that the frequency of occurrence of a particular exonic sequence in the combinatorial library may also be influenced by, for example, varying the concentration of that exon relative to the other exons present, or altering the flanking intronic sequences of that exon to either diminish or enhance its trans-splicing ability relative to the other exons being admixed.

However, the present trans-splicing method can also be utilized for ordered gene assembly, and carried out in much the same fashion as automated oligoucleotide or polypeptide synthesis. FIG. 7 describes schematically the use of resin-bound combinatorial units in the ordered synthesis of a gene. In the illustrated example, mammalian pre-mRNA introns are used to flank the exon sequences, and splicing is catalyzed by addition of splicing extract isolated from mammalian cells. The steps outlined can be carried out manually, but are amenable to automation. The 5' terminal exon sequence (shown as exon 1 in FIG. 7) is directly followed by a 5' portion of an intron that begins with a 5' splice-site consensus sequence, but does not include the branch acceptor sequence. The flanking intron fragment further includes an added nucleotide sequence, labeled "A" in the diagram, at the 3' end of the downstream flanking intron fragment. The 5' end of this terminal combinatorial unit is covalently linked to a solid support.

In the illustrated scheme, exon 2 is covalently joined to exon 1 by trans-splicing. The internal shuffling unit that contains exon 2 is flanked at both ends by intronic fragments. Downstream of exon 2 are intron sequences similar to those downstream of exon 1, with the exception that in place of sequence A the intronic fragment of exon 2 has an added sequence B that is unique, relative to sequence A. Exon 2 is also preceded by a sequence complementary to A (designated A'), followed by the nuclear pre-mRNA intron sequences that were not included downstream of exon 1, including the branch acceptor sequence and 3' splice-site consensus sequence AG.

To accomplish the trans-splicing reaction, the shuffling units are allowed to anneal by hydrogen bonding between the complementary intronic sequences (e.g. A and A'). Then, trans-splicing is catalyzed by the addition of a splicing extract which contains the appropriate snRNPs and other essential splicing factors. The Y-branched intron that is generated, and any other by-products of the reaction, are washed away, and a ligated exon 1 and 2 remain bound to the resin. A second internal shuffling unit is added. As shown in FIG. 7, the exon (exon 3) has flanking intronic fragments which include a sequence B' in the upstream fragment and a sequence A in the downstream fragment. The nucleotide sequence B' is unique relative to sequence A', and is complementary to sequence B. As above, the RNA is allowed to anneal through the B:B' sequences, splicing of the intervening sequences is catalyzed by the addition of extract, and reaction by-products other than the resin bound exons are washed away. While FIG. 7 depicts a non-random assembly of a gene, it is understood that semi-random assembly can also be carried out, such as would occur, for example, when exon 3 is substituted with a variegated population of exons combinatorial units.

This procedure can be continued with other exons, and may be terminated by ligation of a 3' terminal shuffling unit that contains an exon (exon 4 in the FIG. 7) with upstream intron sequence (and either the A' or B' sequence, as appropriate), but lacking any downstream intron sequences. After the 3' terminal exon is added, the assembled gene can be cleaved from the solid support, reverse transcribed, and the cDNA amplified by PCR and cloned into a plasmid by

standard methods.

The domain shuffling experiments described to yield novel protein coding genes can also be used to create new ribozymes. FIG. 11 depicts one example of how group I intron sequences can be used to shuffle group II intron domains. In the illustrative embodiment, the group II intron consists of 6 domains and is flanked by exons (E5 and E3); in this instance, E5 is shown to include a T7 promoter. The six shuffling competent constructs diagrammed in the figure can be made either by standard site directed mutagenesis and cloning or by the reversal of splicing. The 5' terminal exon is followed by sequences from the T4 td intron, beginning with the first nucleotide of the intron and including the internal guide sequence, and continuing through the 5' half of the P6a stem (i.e. including half of L6). The last nucleotide of the exon is a U. The internal guide sequence of the intron is changed by site directed mutagenesis so that it is complementary to the last 6 nt of the exon. This will allow the P1 stem to form. The U at the end of the exon is based paired with a G in the internal guide sequence. The 3' terminal "exon", in this case, consists of group II intron domain 6 plus E3. The 3' terminal exon is preceded by the T4 td intron, beginning with the 3' half of P6a and continuing through to the end of the intron. The last nucleotide of the intron is followed by the first nucleotide of group II intron domain 6. The internal exons each consist of a group II intron domain but, in contrast to the terminal exons, each internal exon is flanked by group I intron sequences on both sides. In each case, the internal guide sequence of the group I intron is changed so as to be complementary to the last 6 nts of the exon and, in each case, the last nucleotide of the exon is a U.

Constructing a library of group II domains flanked by group I intronic sequence allows new group II ribozymes to be assembled from these units by random exon shuffling using conditions that allow for efficient trans-splicing of "exons" flanked by these group I intron sequences. For instance, if only one E5:d1 and d6:E3 are used, but a variegated population of d2-d5, the assembled genes will all have the same 5' and 3' terminal exons, but will have different arrangements and numbers of internal exons. An E3 specific primer plus reverse transcriptase can be used to make cDNA of the library of recombined transcript. T7 and E3 specific primers can be used to amplify the assembled genes by PCR, and RNA transcripts of the assembled gene can be generated using T7 polymerase. The RNA can be incubated under self splicing conditions appropriate for group II splicing. Molecules that are capable of self splicing will yield intron lariats that migrate anomalously slow on denaturing polyacrylamide gels. The lariats can be gel purified and represent active ribozymes. The isolated lariats can be specifically debranched with a HeLa debranching activity. Reverse transcription and PCR can be used to make and amplify cDNA copies of the ribozymes. The primers used for the PCR amplification will include exon sequences so that each amplified intron will be flanked by a 5' and a 3' exon. The last 6 nt of the 5' exon will be complementary to EBS 1. The amplified DNA can be cloned into a plasmid vector and individual interesting variants isolated and studied in detail.

FIG. 12 illustrates an "exon-trap" assay for identifying exons (in the traditional use of the term) from genomic DNA, utilizing trans-splicing mediated by discontinuous nuclear pre-mRNA intron fragments. One advantage of this method is that the DNA does not have to be cloned prior to using the method. In contrast to prior techniques, the starting material of the exon-trap assay could ultimately be total human genomic DNA. In addition, the present method described herein is an in vitro method, and can be easily automated.

In the first step, purified RNA polymerase II is used to transcribe the target DNA. In the absence of the basal transcription factors, Pol II will randomly transcribe DNA (Lewis et al. (1982) Enzymes 15: 109-153). FIG. 12 shows that some of these transcripts will contain individual exons flanked by intron sequences. Since human exons are small, typically less than 300 nt (Hawkins et al. (1988) Nucleic Acids Res. 16, 9893-9908) and introns are large (up to 200,000 nt, Maniatis, T. (1991) Science 251, 33-34) most transcripts will contain either zero or one exon. In the illustrative embodiment, a spliced leader RNA of, for instance, trypanosome or nematode (Agabian (1990) Cell 61, 1157-1160), is covalently linked to a solid support by its 5' end. The RNA generated by random transcription of the genomic DNA is mixed with the immobilized spliced leader and splicing is catalyzed using splicing extract. The resin is then washed to remove unwanted reaction products, such as unreacted RNA and the splicing extract.

Furthermore, in a subsequent step, an in vitro polyadenlyation reaction (for example, Ryner et al. (1989) Mol. Cell. Biol., 9, 4229-4238) can be carried out which adds oligo-A (up to a length of 300 nt) to the 3' end of the RNA. FIG. 12 shows that an RNA transcript, generated by in vitro transcription of a plasmid having an oligo T stretch, followed by the 3' portion of an intron (including the branch acceptor site and the AG dinucleotide), followed by an exon, can be annealed to the immobilized polyadenylated RNA by hydrogen bonding between the poly-A and poly-T sequences. In vitro trans-splicing, catalyzed by splicing extract, will join the known 3' exon to the "trapped" exon. The RNA can then be stripped from the column, copied to DNA by reverse transcriptase and amplified by PCR using primers to the 5' leader and known 3' exon. The amplified DNA that contains a trapped exon will be larger than the side product that results from splicing of the spliced leader exon to the known 3' exon. Thus, the amplified DNA that contains trapped exons can be selected by size.

Moreover, a "capping" reaction can be done to eliminate products that do not contain a trapped exon. After the step of mixing genomically derived RNA with the immobilized exon, a "capping RNA", with a 3' splice site and a 3' exon, can be added and splicing catalyzed by the addition of splicing extract. The 3' exon of the capping RNA is different from the 3' exon of the RNA shown with the oligo-T stretch. The capping RNA is one which will trans-splice very efficiently to any spliced leader RNA which has not already participated in a splicing reaction; but, will splice less efficiently to immobilized RNAs that have a trapped exon ligated to them as the capping RNA lacks a poly-T sequence to anneal to the trapped exon. Therefore, after the capping reaction, the step shown for splicing of the oligoT containing construct will result, primarily, in the generation of the desired (leader/trapped exon/known exon) product and not in the generation of the unwanted (5' leader/3' known exon) product.

III. Cis-splicing Combination of Exons

In yet another embodiment, the combinatorial method can be carried out in a manner that utilizes the flanking intronic sequences in a cis-splicing reaction to generate a combinatorial gene library. As illustrated schematically in FIG. 10, the actual combinatorial event takes place at the DNA level through annealing of complementary sequences within the intron encoding fragments. Briefly, complementary DNA strands are synthesized which correspond to the exonic sequences and flanking intron fragments. As used herein, the term (+) strand refers to the single-stranded

DNA that is of the same polarity as a trans-splicing RNA transcript. That is, intronic sequences flanking the 5' end of the exon represent a 3' fragment of an intron. Likewise, the term (-) strand refers to the single stranded DNA which is complementary to the (+) strand (e.g. of opposite polarity).

The 5' and 3' ends of each of the (+) and (-) strands are complementary and can therefore mediate concatenation of single-stranded DNA fragments to one and other through basepairing. In the exemplary illustration of FIG. 10, the exon sequences are flanked by group II domains IV-VI at one end, and domains I-IV at the other. A library of combinatorial units representative of a number of different exons is generated, such as by PCR or digestion of double-stranded plasmid DNA, to include both (+) and (-) strands. The units are combined under denaturing conditions, and then renatured. Upon renaturation, the sequences corresponding to domain IV at the 3' end of one (+) strand unit can anneal with the complementary domain IV sequences at the 3' end of a (-) strand unit, resulting in concatenation of combinatorial units (see FIG. 10).

Double-stranded DNA can be generated from the concatenated single-stranded units by incubating with a DNA polymerase, dNTPs, and DNA ligase; and the resulting combinatorial genes subsequently cloned into an expression vector. In one instance, 5' terminal and 3' terminal combinatorial units can be used and the double-stranded genes can be amplified using PCR anchors which correspond to sequences in each of the two terminal units. The PCR primers can further be used to add restriction endonuclease cleavage sites which allow the amplified products to be conveniently ligated into the backbone of an expression vector. Upon transcription of the combinatorial gene, the intronic RNA sequences will drive ligation of the exonic sequences to produce an intron-less transcript.

While FIG. 10 demonstrates one embodiment which utilizes group II introns, the combinatorial process can be carried out in similar fashion using either group I intron sequences or nuclear pre-mRNA intron sequences.

IV. Circular RNA transcripts

In addition to generating combinatorial gene libraries, the trans-splicing exon constructs of the present invention have a number of other significant uses. For instance, the present trans-splicing constructs can be used to produce circular RNA molecules. In particular, exon constructs flanked by either group II or nuclear pre-mRNA fragments can, under conditions which facilitate exon ligation by trans-splicing of the flanking intron sequences, drive the manufacture of circularly permuted exonic sequences in which the 5' and 3' ends of the same exon are covalently linked via a phosphodiester bond.

Circular RNA moieties generated in the present invention can have several advantages over the equivalent "linear" constructs. For example, the lack of a free 5' or 3' end may render the molecule less susceptible to degradation by cellular nucleases. Such a characteristic can be especially beneficial, for instance, in the use of ribozymes in vivo, as might be involved in a particular gene therapy. In the instance of generating ribozymes, the "exonic" sequences circularized are not true exons in the sense that they encode proteins, rather, the circularized sequences are themselves intronic in origin, and flanked by other trans-acting intron fragments.

However, the circularization of mature messenger-RNA transcripts can also be beneficial, by conferring increased stability as described above, as well as potentially increasing the level of protein translation from the transcript. To illustrate, a ribosome which has completed translation of a protein from the present circular transcript may continue to track around the transcript without dissociating from it, and hence renew synthesis of another protein. Alternatively, the ribosome may dissociate after translation is completed but, by design of the circular transcript, will disengage the transcript proximate to the start site and thereby provide an increased probability that the ribosome will rebind the transcript and repeat translation. Either scenario can provide a greater level of protein translation from the circular transcript relative to the equivalent linear transcript.

FIGS. 13A and B depict two examples of intron fragment constructs, designated (IVS5,6)-exon-(IVS1-3), and (3'-half-IVS)-exon-(5'-half-IVS), which, in addition to being capable of driving trans-splicing between heterologous exons as described above, can also be used to generate circular RNA transcripts. The (IVS5,6)-exon-(IVS 1-3) transcript comprises the group II intron domains V and VI at the 5' end of the exon, and domains I-III at the 3' end of the exon. The (3'-half-IVS)-exon-(5'half-IVS) is a similar construct, but replaces the group II domains V-VI and I-III with fragments corresponding to the 3'-half and 5'-half of a nuclear pre-mRNA intron. As described in Examples 1 and 2 below, each of these transcripts can be shown to drive intramolecular ligation of the exon's 5' and 3' end to form circular exons.

Furthermore, as set forth in Example 2, a preferred embodiment of an exon construct using mammalian pre-mRNA intron sequences to generate circular transcripts provides an added structural element that brings together the 5' and 3' ends of the flanking pre-mRNA intron fragments. The addition of such structural elements has been demonstrated to greatly improve the efficiency of the intramolecular splicing reaction. For example, the ends of the intronic fragments can be non-covalently linked as shown in FIG. 13B, by hydrogen bonding between complementary sequences. Alternatively, the ends of the nuclear pre-mRNA intron fragments can be covalently closed. In an illustrative embodiment, FIG. 14 shows how group II intronic fragments can be utilized to covalently join the ends of the nuclear pre-mRNA transcripts having flanking nuclear pre-mRNA intron fragments, which subsequently drive ligation of the 5' and 3' end of the exonic sequences.

In yet another embodiment, the intronic ends can be brought together by a nucleic acid "bridge" which involves hydrogen bonding between the intronic fragments flanking the exon and a second discrete nucleic acid moiety. As illustrated in FIGS. 15A-C, such nucleic acid bridges can be formed a number of ways. Each of the splicing bridges shown differ from each other in either the orientation of the bridge oligonucleotide when basepaired to the flanking intron fragments, in the size of the bridging oligonucleotide, or both.

For instance, the bridge oligonucleotide shown in FIG. 15A base-pairs in an orientation which can result in a stem-structure similar to the (3' IVS-half)-exon-(5' IVS-half) construct depicted in FIG. 13B. Moreover, when a bridge similar to one shown in FIG. 15C is used, and the 5' and 3' ends of the flanking introns base-pair some distance apart in the linear sequence of the bridge, the bridge oligonucleotide may itself comprise the branch acceptor site. For example, the bridge

oligonucleotide can be an RNA transcript comprising the yeast branch site consensus sequence UACUAAC in a portion of the bridge sequence which does not base-pair with the intronic fragments of the exon construct.

Oligonucleotide bridges useful in driving the circularization of exon transcripts can also be used to direct alternative splicing by "exon skipping", which may be useful, for example, in disrupting expression of a particular protein. As shown in FIG. 15D, the splicing of exons 1 and 3 to each other can be the result of an oligonucleotide which loops out exon 2, effectively bringing together two complementary halves of the intronic sequences flanking exons 1 and 3. As shown in FIG. 15D, exon 2 can, in fact, be spliced into a circular RNA.

Carrying the bridging nucleotide one step further, FIG. 16 illustrates the use of an exon construct useful in mediating the alternate splicing of an exon through a trans-splicing-like mechanism. For instance, a wild-type exon can be trans-spliced into an mRNA transcript so as to replace an exon in which a mutation has arisen. The wild-type exon construct comprises flanking intronic sequences which include sequences complementary to a portion of the continuous introns which connect exons 1, mutant exon 2, and exon 3. Thus, through a trans-splicing event as described above, some of the resulting mature mRNA transcripts will include the wild-type exon 2.

EXAMPLE 1

Group II Introns can Mediate Circularization of Exonic Sequences

The (IVS 5,6)-exon-(IVS 1-3) RNA transcript, shown in FIG. 13A, was synthesized from plasmid pINV1 (Seg. ID No. 1). The intronic sequences correspond to the half molecules generated by interruption of the 5 g intron of the yeast mitochondrial oxi3 gene in domain IV; and the exonic sequences are the exon sequences E5 and E3 which are naturally disposed at the 5' and 3' ends of the 5 g intron, respectively. To construct pINV1, the Sac I-Hind III fragment of pJDI5'-75 (Jarrell et al. (1988) Mol. Cell Biol. 8:2361-2366) was isolated and the Hind III site was filled in with Klenow fragment. This DNA was ligated to pJDI3'-673 (Jarrell et al., supra) that had been cleaved with Sac I and Sma I. The RNA splicing substrates were made by in vitro transcription using T7 RNA polymearse.

Transcription, RNA purification, and splicing reactions were as described (Jarrell et al., supra). The E5-specific oligodexoynucleotide (5'-GTAGGATTAGATGCAGATACTAGAGC-3' (SEQ ID NO. 6)) is identical to 26 nucleotides of the E5 region of the (IVS 5,6)exon(IVS 1-3) RNA. The E3-specific oligonucleotide (5'-GAGGACTTCAATAGTAGTATCCTGC-3' (SEQ ID NO. 7)) is homologous to 25 nt of the E3 region.

To purify E3,E5(C), described below, for the reverse transcription reaction, a standard 100-.mu.l transcription was done, with pINV1 as a template. The (IVS 5,6)E3,E5(IVS 1-3) RNA was concentrated by ethanol precipitation and was then incubated under the (NH.sub.4).sub.2 SO.sub.4 splicing conditions for 1 hr. The E3,E5(C) RNA was gel purified and dissolved in 30 .mu.l of water. A 9-.mu.l annealing reaction mixture was incubated at 65.degree. C. for 3 min and then placed on ice. The annealing reacting mixture included 1 .mu.l of the E3,E5(C) RNA plus 100 ng of the E3-specific oligonucleotide. As a control, an identical annealing reaction was

done, except E3,E5(C) was not added. A buffer (4 .mu.l) consisting of 0.25M Tris-HCl (pH 8.5), 0.25M KCl, 0.05 M dithiothreitol, and 0.05M MgCI.sub.2 was added to both annealing reaction mixtures. Deoxynucleoside triphosphates were each added to a final concentration of 5 mM, followed by 40 units of RNasin (Promgea) and 22 units of reverse transcriptase (Seikagaku America, Rockville, Md.). The final volume was adjusted to 20 .mu.l with water. The mixture was incubated at 42.degree. C. for 90 min.

Two polymerase chain reaction (PCR) experiments were done using as templates either 1 .mu.l of the reverse transcription mixture that included E3,E5(C) or 1 .mu.l of the control reverse transcription mixture, which lacked E3,E5(C). The PCRs were performed as described (14) and were continued for 25 cycles. The E3- and E5-specific oligonucleotides, 300 ng each, were used as PCR primers. DNA sequencing was done with Sequence (United States Biochemical) according to the protocol provided by the manufacturer.

Group II intron excision can occur by transesterification (splicing) or by site-specific hydrolysis (cleavage). The former reaction is stimulated by (NH.sub.4).sub.2 SO.sub.4, the control RNAs, E5(IVS 1-3) plus (IVS5,6)E3, trans-spliced to yield spliced exon S(E5-E3) and a Y-branched intron [IVS(Y)]. Coincubation in the presence of KCl yielded free exons (E5 and E3) and a linear intron (IVS 1-3) as major products.

The (IVS 5,6)exon(IVS 1-3) precursor was also reactive. Most for the products could be identified based on their comigration with products of the control trans-reaction. In the presence of (NH.sub.4).sub.2 SO.sub.4, the IVS(Y) and some linear intron were liberated; several novel products were also generated. Among these was an RNA (E3,E5) the expected size of the linear excised exons (591 nt). A slower migrating RNA [E3,E5(C)] was also observed. At short times of incubation (1 min) E3,E5(C) and IVS(Y) were the predominant products. In contrast, E3,E5 did not accumulate to significant levels before 60 min, indicating that it was not an early product of the reaction. Analysis of E3,E5(C) demonstrated that is was circular spliced exons. E3,E5(C) accumulated in the presence of (NH.sub.4).sub.2 SO.sub.4 but not in the presence of KCl. This was significant, given that spliced exons (E3-E5) are not only product of cis or trans splicing that accumulates in the presence of (NH.sub.4).sub.2 SO.sub.4 but not in the presence of KCl. Thus, it was likely that E3,E5(C) resulted from splicing rather than hydrolysis.

E3,E5(C) and E3,E5 were purified and analyzed by denaturing gel electrophoresis. During the purification process some E3,E5(C) was converted to a faster migrating species that comigrated with E3,E5. The extent of conversion of E3,E5(C) to the faster migrating species was increased by incubation with the group II intron under conditions that promote site-specific hydrolysis of the spliced exons. These observations are consistent with E3,E5(C) being a circular RNA that can be broken by hydrolysis to yield (linear) E3,E5.

To demonstrate that E3,E5(C) contains spliced exons, a cDNA copy of purified E3,E5(C) RNA was made by reverse transcription. The reverse transcription was primed with an oligonucleotide homologous to 25 nt of E3. If E3,E5(C) is accurately spliced circular exons, its length is 591 nt. Reverse transcription of this circular RNA would yield cDNAs of variable lengths; in particular, multiple rounds of complete reverse transcription of the circular template would generate cDNAs that are >591 nt long. A sample of the reverse transcription reaction mixture was used as a

template in a PCR. The E3-specific oligonucleotide and an oligonucleotide homologous to 26 nt of the E5 sequence of the expected cDNA were used as primers. If E3,E5(C) is the product of a splicing reaction, it will contain both E3 and E5 sequences and will yield amplification products in this PCR reaction. Analysis of the PCR products revealed that the major amplification product is the size expected [313 base pairs (bp)] for a PCR product derived from spliced exons. This product was not seen in a control PCR reaction. Two additional PCR products of about 900 bp and 1500 bp were also observed. Amplification of longer cDNAs generated by multiple rounds of reverse transcription of the circular E3,E5(C) template would yield a set PCR products each an integral multiple of 591 bp longer than the 313 bp indicating that the 900 bp and 1500 bp observed products were likely generated in this manner.

The 313-bp PCR product was purified and cloned into a plasmid vector. The nucleotide sequence of each of four independently isolated clones was determined by the dideoxy sequencing method, using the E3-specific oligonucleotide as a primer. The sequence showed that the PCR product contained both E5 and E3 sequences that were joined by accurate splicing.

EXAMPLE 2

Mammalian Nuclear Pre-mRNA Introns can Mediate circularization of exonic sequences

The BGINV plasmid (SEQ ID NO. 2) was derived from plasmid HBT7. HBT7 has the first intron of the human .beta.-globin gene, flanked by .beta.-globin exon 1 and 2 sequences, cloned into the psp73 vector. To construct BGINV, HBT7 was cut at the unique BbvI site in the intron and at the unique BamHI site, downstream of Exon 2. The ends were made blunt with klenow fragment. The DNA was diluted and ligase was added. A clone was isolated (BGUS) that had exon 1 and intron sequence, up to the filled BbvII site. In a separate experiment, HBT7 was cut with HindIII and BbvI, the ends were filled in, and the DNA was diluted and ligated. A clone was isolated BGDS, that had intron sequence, beginning with the filled BbvI site, followed by exon 2 sequences. BGDS was cut with XhoI and SmaI and the fragment containing the intron and exon 2 sequences was gel purified. This DNA was ligated into BGUS that had been cleaved with XhoI and PvuII, to yield BGINV. The inverse-.beta.-globin RNA can be transcribed from this plasmid in vitro using T7 polymerase.

BGINV was cut with EcoRI and RNA was transcribed in vitro using T7 polymerase. In vitro splicing reaction were done as described in Hannon et al. (Hannon et al. (1990) Cell, 61, 1247-1255), except mammalian extract was used. The extract was prepared by the method of Dignam et al. (Dignam et al. (1983) Nucl. Acids Res. 11, 1475-1489). Splicing extract is also commercially available (Promega, cat.# E3980). Spliced products were separated by polyacrylamide gel electrophoresis and visualized by autoradiography.

The transcription reaction that generated the RNA that was used to create the circular precursor included GMP (final concentration, 0.8 mM); this was to ensure that some of the RNA transcripts initiated with GMP, instead of GTP, since a 5' phosphate is a substrate for ligase (while a 5' triphosphate is not). The transcript was purified from a polyacrylamide gel. Circular precursors were generated using a DNA oligonucleotide (5'CGAGGCCGGTCTCCCAATTCGAGCTCGGTAC (SEQ ID NO.8)) to bring the ends of the

RNA together, followed by the addition of DNA ligase to covalently join the ends (Moore et al. (1992) Science 256, 992-997). The circular precursor was purified from a polyacrylamide gel. In vitro splicing reactions were done as described above.

The circular exon product was observed and characterized. This RNA was gel purified and a cDNA copy generated using the CIR-1 primer (5'GAGTGGACAGATCCCCAAAGGACTC (SEQ ID NO. 9)) which is specific to exon 2 sequences. The cDNA was amplified by PCR using the CIR-1 and CIR-2 (5'-GTGATGGCCTGGCTCACCTGGACAA (SEQ ID NO. 10)) oligonucleotides as primers. A 145 nt product was observed. This amplification product is the expected size of a product generated from circular spliced exons.

The branched intermediate (generated by the first step of the reaction) was also observed and characterized. It was gel purified and treated with HeLa debranching enzyme (Ruskin et al. (1985) Science 229, 135-140). This treatment increased the rate of migration of the RNA through a denaturing polyacrylamide gel such that it migrated as a 553 nt RNA, consistent with the assignment of the product as the lariat intermediate.

V. Reagents for Molecular Biology

Molecular cloning of DNA currently relies heavily on restriction enzymes and DNA ligase to specifically cut and join molecules. The reverse-splicing introns or "ribozymes" of the present invention can fulfill these two functions; they can both cut and join RNA molecules, and thus can serve as useful tools for nucleic acid manipulation. In similar fashion to the activation of an exon by addition of flanking intronic fragments through the reversal of splicing the recombinant RNA technology described herein involves attacking a target RNA molecule with an intronic molecule and, by the reversal of splicing, cleaving the target into two pieces while simultaneously joining specific intron sequences to the cleaved ends of the target molecule. The newly formed exon construction can be purified, and appropriate exons ligated to each other through trans-splicing mediated by the intronic fragments. Alternatively, these recombinant RNA molecules can be cloned into a plasmid, and fresh RNA transcripts generated from these plasmids, with these second generation transcript being used in a trans-splicing reaction. Thus, cleavage and ligation functions similar to those provided by restriction enzymes and ligase can be provided by RNA trans-splicing.

DNA restriction enzymes and DNA ligase are so routinely used for nucleic acid manipulation that the limitations of these reagents are seldom considered. Restriction enzymes typically recognize and cleave specific DNA sequences that are 4 to 6 basepairs in length. Although there are theoretically 4,096 different possible restriction enzymes that recognize 6 basepair sequences, only 78 such enzymes with distinct specificities are commercially available. One reason that most possible specificities are unavailable is that it is not feasible to engineer the sequence specificity of a restriction enzyme. Instead, micro-organisms must be identified that naturally produce enzymes with novel specificities. Often, it is difficult to obtain large quantities of pure active enzyme from these natural sources, leading to the second limitation, which is that restriction enzymes are often impure and the enzyme concentration is low. A third limitation is that certain classes of DNA sequences are not recognized by any known restriction enzyme. For example, there are no known enzymes that recognize sequences comprising only of A and C

nucleotides, such as 5'-AACCAA. A fourth limitation is that DNA ligase only joins DNA molecules with compatible ends, making it often necessary to fill in or degrade 5' or 3' overhangs on DNA molecules before they can be joined by ligation. Finally, yet another limitation is that DNA ligation reactions are often not directional, leading to the generation of recombinant clones with inserts in the wrong orientation.

In contrast, the advantages of the present Y-branched ribozymes is system are that potentially any 3-16 nt sequence can be specifically targeted. Accordingly, whereas restriction enzymes are much more limited, recognizing only a small subset of, for example, the 4096 possible 6 nt sequences present in DNA, the subject ribozymes can be generated for each of the 4,096 different sequences. Indeed, under appropriate reaction conditions, the efficiency of the reaction can be greatly influenced by the EBS2-IBS2 interaction such that the specificity of the ribozyme is effectively 12-16 basepairs. Consequently, in the instance of the ribozyme which recognizes a specific 12 nucleotide target sequence, over 16 million different specificities are possible and can be accessed by the present invention.

Moreover, in contrast to restriction enzymes which typically require palindromic sequences that may introduce ambiguity into the orientation of DNA sequences inserted at a restriction endonuclease cleavage site the subject ribozymes can be orientation specific. In addition, once an RNA is followed by, or preceded by, the correct intron sequences, any upstream molecule can be joined to any downstream molecule (see, for instance, appended Examples 9 and 10). In contrast, when molecular cloning is done with restriction enzymes, only molecules with compatible ends can be joined; for example, a molecule with Eco RI ends cannot be joined to a molecule with Hind III ends without first filling in the ends. Furthermore, molecules that are joined by trans-splicing are "seamless". That is, recognition sites do not have to be engineered into the target molecules in order to cleave and ligate the target molecule. Instead, the ribozyme is engineered to match the target. For instance, a library of reverse-splicing ribozymes can be generated to comprise every possible 6 nucleotide combination by manipulating intron sequences which interact with the "exon" target (e.g. the IBS1 for group II, and the IGS for group I). Thus, sequences can be precisely joined without adding, deleting or changing any nucleotides. Finally, for the autocatalytic introns, no enzymes need be added in order to catalyze the forward or reverse reactions. Instead, the RNAs are incubated together in a simple salt solution and other appropriate ions and the recombinant molecules are generated.

Accordingly, one aspect of the invention pertains to a preparation of a reverse-splicing intron which comprises two or more fragments of autocatalytic introns and catalyzes integration of at least a portion of the reverse-splicing contruct into a substrate ribonucleic acid by a reverse-splicing reaction. For example, the autocatalytic intron fragments can be derived from one or more group II introns, and preferably are derived with exon binding site which have been altered by recombinant mutagenesis. In another illustrative embodiment, the autocatalytic intron fragments are derived from group I introns. Again, the specificity of the intron is preferably altered by recombinant mutagenesis of the internal guide sequence of group I intron fragments.

In another embodiment, as is apparent from the description throughout the present application, where the reverse-splicing intron is derived from a group II intron, it may comprise a first segment having a 5' portion of a group II intron, which 5' portion includes an exon binding site;

and a second segment comprising a 3' portion of a group II intron, which 3' portion includes a domain V motif, a branch site acceptor forming a phosphodiester bond with the 5' end of the first segment, and a nucleophilic group at the 3' end of the second segment for transesterifying a phosphodiester bond of a ribonucleic acid. By this arrangement, the first and second segments together form an autocatalytic Y-branched intron which catalyzes integration of at least the first segment of the reverse-splicing intron into a substrate ribonucleic acid by a reverse-splicing reaction.

In an exemplary embodiment, the 5' portion of the group II intron includes intron domains V and VI, and the 3' portion of the group II intron includes intron domains I--III. Moreover, it will be understood that the reverse-splicing construct can be a Y-branched lariat form of the group II intron, e.g., the first and second segments are contiguous via a covalent bond other than the phosphodiester bond formed with said branch site acceptor, or can be in the form of a Y-branched discontinous intron, e.g., the first and second segments are covalently attached only at the branch site acceptor. For instance,the reverse-splicing intron can be represented by the general formula: ##STR1## wherein (IVS 1-3) represents a 5' portion of a group II intron,

(IVS5,6) represents a 3' portion of a group II intron, which portion includes a branch site acceptor,

2'-5' represents a phosphodiester bond formed between a branch site acceptor of (IVS5,6) and the 5' end of (IVS 1-3), and

A, if present, represents a phosphodiester bond between a 3' end of (IVS 1-3) and a 5' end of (IVS5,6), wherein (IVS1-3) and (IVS5,6) together form an autocatalytic Y-branched intron which catalyzes integration of at least the (IVS1-3) fragment, if discontinuous with (IVS5,6), into a substrate ribonucleic acid by a reverse-splicing reaction.

In preferred embodiments, the exon binding site, e.g., EBS1 and/or EBS2 is altered (or deleted in certain instances) by recombinant mutagenesis. As a result, the exon binding site can be chosen to provide specific integration into a substrate ribonucleic acid at a selected intron binding site, such that the effective exon binding sequence can be from 3-16 nucleotides in length.

In yet another preferred embodiment, the reverse splicing intron is provided as a substantially pure preparation. By "substantially pure" it is meant that the construct has been isolated from, or otherwise substantially free of other polynucleotides, especially exonic sequences, normally associated with the intron. The term "substantially pure" or "substantially pure or purified preparations" are defined as encompassing preparations of the reversing splicing introns having less than 20% (by dry weight) contaminating protein or polynucleotides, and preferably having less than 5% contaminating protein or polynucleotides. By "purified", it is meant, when referring to a nucleic acid construct of the present invention, that the indicated molecule is present in the substantial absence of other biological macromolecules, such as other proteins or polynucleotides. The term "purified" as used herein preferably means at least 80% by dry weight, more preferably in the range of 95-99% by weight, and most preferably at least 99.8% by weight, of biological macromolecules of the same type present (but water, buffers, and other small molecules, especially molecules having a molecular weight of less than 3000, can be present).

The term "pure" as used herein preferably has the same numerical limits as "purified" immediately above. "Isolated" and "purified" do not encompass either natural materials in their native state or natural materials that have been separated into components (e.g., in an acrylamide gel) but not obtained either as pure (e.g. lacking contaminating proteins or polynucleotides, or chromatography reagents such as denaturing agents and polymers, e.g. acrylamide or agarose) substances or solutions.

To further illustrate, a group II intron or portion thereof can be used to specifically cut and join RNA molecules. As described above, the group II intron splicing reaction is reversible. If an intron lariat, a product of the forward reaction, is incubated with spliced exons at high RNA concentration under the reaction conditions used for the forward reaction, the intron specifically inserts into the spliced exons, thus regenerating the precursor RNA (see FIG. 1). Likewise, as illustrated in FIG. 6, a Y-branched form of the intron, generated for example by an inverse splicing reaction, can also insert into spliced exons. When a Y-branched intron, such as the illustrated (IVS5,6)2'-5'(IVS1-3) lariat, is used in a reverse-splicing reaction, the exon target is cleaved into two pieces. The upstream piece becomes joined to intron domains 1-3 and the downstream piece becomes joined to intron domains 5 and 6.

The 3-8 nucloetide EBS 1 site on the ribozyme is the primary determinant of the specificity of the reverse reaction for group II introns. In the reverse reaction, EBS 1 selects the site of integration by hydrogen bonding to it. The intron is subsequently inserted just downstream of this target sequence. By changing the nucleotide sequence of EBS 1, the ribozyme can be targeted to insert downstream of any specific 3-8 nt sequence. Moreover, the manipulation of the EBS 2:IBS 2 interactions can also influence the efficiency of splicing and provide even greater specificity to the insertion site (e.g. by expanding the recognition sequence to, for example, 10-16 nucleotides). Likewise, manipulation of the IGS, and other secondary intron:exon contacts analogous to EBS2, the specificity of a group I reverse splicing ribozyme, such as (IVS P1-P6.5)(IVS P6.5-P10) can be controlled.

FIG. 18 depicts a further embodiment illustrating how a reverse-splicing ribozyme, such as the group II lariat IVS, can also be used to cleave and ligate target RNA molecules. The site directed mutagenesis is the same as described above (the EBS 1 and IBS 1 sequences are changed). The lariat ribozyme is generated by the forward reaction. The reverse reaction yields a single molecule with the intron specifically inserted in it. A cDNA copy is made by reverse transcriptase. Two different sets of PCR primers are used to amplify either the upstream portion of the interrupted target molecule, plus intron domains I-III-3 or to amplify domains V and VI and the downstream portion of the target molecule. Each of these amplified DNAs can be cloned into a plasmid to generate the same two constructs shown in FIG. 17.

In another illustrative embodiment, FIG. 19 depicts a method by which the present trans-splicing constructs can be used to manipulate nucleic acid sequences into a plasmid such as a cloning or expression vector. In such a scheme, the plasmid sequence is itself an exon being flanked at each end by intronic fragments capable of mediating a trans-splicing reaction. For example, as shown in FIG. 19, the plasmid can be generated as an RNA transcript comprising the backbone sequences of the plasmid, flanked at the 5' end with the group II domains V and VI, and at the 3' end with the group II and domains I-III. To generate such a transcript, a pre-plasmid can be

utilized in which the 5' and 3' flanking sequences are joined with an intervening sequence including a T.sub.7 RNA promoter sequences and endonuclease cleavage site. The plasmid is linearized by cleavage at the endonucleasesensitive site, and the linearized plasmid transcribed to RNA using standard techniques.

The nucleic acid sequences to be cloned into the plasmid is generated to similarly include flanking group II intron fragments. Mixing the two transcripts under trans-splicing conditions will therefore result in ligation of the nucleic acid of interest into the plasmid, in the appropriate orientation and at the correct site. Such a method is particularly amenable to the closing of the above-described combinatorial gene libraries into replicable expression vectors. Furthermore, this trans-splicing technique of sub-cloning can be used effectively in random mutagenesis applications. For instance, the nucleic acid of interest can be first treated with actinic acid such that a discrete number of base modifications occur, and then ligated into the plasmid.

Another aspect of the invention pertains to a kit for generating a Y-branched ribozyme of a particular specificity. In general, the kit can feature an expression vector containing a gene, whose expression product will give rise to a Y-branched ribozyme of the present invention. For instance, the transcript illustrated by FIG. 17, and described in further detail in Example 3 below, can be used to generate the Y-branch ribozymes of the present invention. For instance, a vector can be constructed containing such a gene having unique restriction enzyme sites immediately 5' to the IBS2 sequence and 3' to the EBS 1 sequence, such that the EBS1 and (optionally) the EBS2 sequences can be altered by insertion of oligonucleotide cassettes. In another embodiment, the restriction sites can be placed immediately flanking the EBS1-EBS2 sites and another set of restriction sites used to flank the IBS1-IBS2 site such that 2 oligonucleotide are used to alter the EBS1 and EBS2 specificities. Where a continuous Y-branched laureate structure is desired a construct as shown in the forward reaction of FIG. 1, e.g., exonl-intron-exon2, can be used and appropriately manipulated to yield a certain specificity for the EBS1 and EBS2 recognition sequences.

Alternatively, for high throughput needs, automated systems can be provided for scale-up production of quantities of the subject inverse-splicing constructs. An exemplary approach for high throughput production in commercial scale synthesis of the subject ribozymes can rely on the fact that three pieces of the IVS(1-3) fragment (which flank the EBS1 and EBS2 sequences) can be identical in each intron, and can consequently be produced in bulk by RNA synthesis procedures. For instance, T7 transcription processes have been scaled up to permit purification of milligram or greater quantaties of RNA. Likewise, the EBS1 and EBS2 sequences can be generated easily by RNA solid phase synthesis. Likewise, two DNA oligonucleotides (1 and 2) can be synthesized by standard automated approaches. Each of the two DNA oligos is homologous to one of the EBS sequences. and to the flanking intron sequences. The mixture is annealed to produce a DNA/RNA duplex, and the nicks in the RNA strand can be sealed using a DNA ligase. The DNA oligonucleotides are than removed by treatment with DNase I. Similar procedures. using DNA/DNA pairs can be carried out in place of the use of restriction sites described above.

Still another embodiment of the present invention pertains to a library of reverse-splicing introns comprising a variegated population of Y-branched group II introns. In preferred embodiments,

the variegated population is characterized as including at least 25 different Y-branched group II introns of unique specificity, more preferably at least 100 different Y-branched group II introns of unique specificity, and even more preferably from 10.sub.3 to 10.sub.6 different Y-branched group II introns of unique specificity.

EXAMPLE 3

Use of group II Y-branched lariats as endonucleases/ligase

FIG. 17 is an exemplary illustration of the use of these reactions to generate recombinant molecules. The last six nucleotides of the (IVS5,6)E3,E5 (IVS1-3) RNA, which was generated by in vitro transcription of pINV1, are ATTTTC. The EBS 1 sequence of the flanking intron fragment is GGAAAT. As described in Example 9 below, inverse splicing of RNA transcribed from pINV1 yields a Y-branched intron with a wild-type EBS 1 sequence (GGAAAT). FIG. 17 shows a 404 nt RNA (TPA S,F) that includes coding information for the signal sequence and growth factor domain of the TPA cDNA clone. This transcript was generated from plasmid TPA-KS+that had been cut with Sty I. The goal was to attack TPA S,F with a Y-branched ribozyme such that the ribozyme inserted downstream of the GTCAAA sequence that is present at the end of the growth factor domain. In order to use pINV1 to generate a Y-branched ribozyme capable of attacking the TPA S,F RNA, the EBS 1 and IBS 1 sequences of pINV1 were changed by site directed mutagenesis. The IBS 1 sequence was changed to GTCAAA (that is, to the same sequence present in the TPA transcript that is to be attacked), and the EBS 1 sequence was changed to TTTGAC in order that it be complementary to the mutated IBS 1 sequence. RNA was transcribed in vitro from this altered plasmid (termed here GrII-SIG) and incubated under splicing conditions to yield the excised Y-branched molecule (SIG-Y). This Y-branched intron is identical to that derived from (IVS5,6)E5,E3,(IVS1-3) in Example 9, except the EBS 1 sequence is TTTGAC. This Y-branched ribozyme was tested for its ability to insert specifically into TAP S,F RNA. As diagrammed in FIG. 17, this RNA was incubated with the 404 nt target RNA under splicing conditions. Specific reversal generates a 1047 nt product that consists of the first 332 nt of the TPA-KS+transcript ligated to intron domains 1-3. This 1047 nt product was gel purified and a cDNA copy was made by reverse transcription. The cDNA was amplified by PCR and cloned into a vector to yield plasmid SIG(IVS1-3). The smaller, 108 nt, product consists of intron domains 5 and 6 ligated to 72 nt of TPA S,F. A cDNA copy of the smaller product can likewise be made by reverse transcription, amplified by PCR, and the amplified product cloned into a vector to yield plasmid (IVS5,6)StyI.

Following the success of the inverse splicing reaction, the role of the IBS2-EBS2 interaction was investigated with respect to efficiency of the reverse splicing reaction. Starting with the construct pINV1, oligonucleotide primer mutagenesis was used to alter the IBS 1 sequence to CTGCTCC and the EBS 1 sequence to GGAGCAG. The EBS2 sequence was changed by oligonucleotide directed mutagenesis to the sequence GGCACA, while the IBS2 sequence was changed to the corresponding TGTGCC, to yield the construct pY7. Surprisingly, the reaction efficiency for the Y-branched ribozymes was several orders of magnitude better in the reversal of splicing reaction involving both the EBS1 and EBS2 interactions, relative to the Y-branch laureate having only a matched EBS1 interaction with the target RNA. Thus, despite the indication in the literature that the EBS2-IBS2 interaction is not essential for Group II intron splicing, the present data would

indicate that this interaction is much more important to the efficiency of the reversal of splicing reaction. Consequently, as set out above, the subject reverse splicing ribozymes can be generated to be solely dependent on the EBS1/IBS1 interaction, e.g., by deletion or mismatch of the EBS2 with the target nucleic acid, or alternatively, can be generated to exploit all or a portion of the EBS2/IBS2interaction by recombinantly engineering the sequence of the EBS2 sequence.

As suggested above, discrete inverse splicing introns can be generated for each of the potentially 4096 different 6 base sequences. However, since G:U base pairing is permissible in RNA, certain EBS1 sequences can give rise to specificity for more than one IBS1 target sequence. For instance, and EBS1 of GGGGGG or UUUUUU could, in the absence of any contribution from the EBS2/IBS2 interaction (e.g., EBS2 deleted) have an affinity for 64 different sequences. Accordingly, a library of 100 ribozymes with specificities that are derived from all pairwise combinations of the following triplets: ACA, CCC, CCA, AAC, AAA, ACC, CAA, CAC, UUU and GGG can be generated which recognize 576 different 6 nucleotide sequences. To illustrate, of this set of ribozymes, 64 would recognize one unique sequence each, 32 of the enzymes would recognize 8 different sequences each, and 4 would recognize 64 different sequences each. This library therefore represents, in specificity, approximately 1/7th of the total possible six nucleotide sequences, and would be expected, on average, have a corresponding recognition sequence (IBS 1) approximately every 42 nucleotides. In contrast, the 78 or so restriction enzymes, on average, have recognition sequences every 320 or so basepairs.

It is clear from this example that potentially any 4-16 nt RNA sequence can be attacked specifically by a Y-branched ribozyme that has been engineered to have the appropriate EBS 1 and (optionally) EBS2 sequence. The target molecule will be split into two pieces. Intron domains 1-3 will be ligated to the upstream piece, while domains 5 and 6 will be ligated to the downstream piece. Following reverse transcription and PCR, these recombinant molecules can each be cloned into a plasmid vector downstream, for example, of the T7 promoter. Synthesis of RNA from the plasmid will yield transcripts capable of trans-splicing. Thus, in the above example, the original 404 nt target RNA could be regenerated by trans splicing. Moreover, it is also true that trans-splicing can be used to join the TPA sequences of SIG(IVS 1-3) to any other RNA that has intron domains 5 and 6 upstream of it. The recombinant RNA molecule generated by such a trans-splicing reaction could be copied into cDNA, amplified by PCR and cloned into a plasmid vector.

VI. Generating Novel Genes and Gene Products

A major goal of the present combinatorial method is to increase the number of novel genes and gene products that can be created by exon shuffling in a reasonable period of time. As described herein, the exon portion of the present splicing constructs can encode a polypeptide derived from a naturally occurring protein, or can be artificial in sequence. The exon portion can also be a nucleic acid sequences of other function, such as a sequence derived from a ribozyme. By accelerated molecular evolution through shuffling of such exons, a far greater population of novel gene products can be generated and screened in a meaningful period of time.

In our embodiment, the field of application of the present combinatorial method is in the generation of novel enzymatic activities, such as proteolytic enzymes. For example,

combinatorial trans-splicing can be used to rapidly generate a library of potential thrombolytic agents by randomly shuffling the domains of several known blood serum proteins. In another embodiment, the trans-splicing technique can be used to generate a library of antibodies from which antibodies of particular affinity for a given antigen can be isolated. As described below, such an application can also be especially useful in grafting CDRs from one variable region to another, as required in the "humanization" of non-human antibodies. Similarly, the present technology can be extended to the immunoglobulin-super family, including the T-cell receptor, etc., to generate novel immulogically active proteins.

In another illustrative embodiment, the present trans-splicing method can be used to generate novel signal-transduction proteins which can subsequently be used to generate cells which have altered responses to certain biological ligands or stimuli. For instance, protein tyrosine kinases play an important role in the control of cell growth and differentiation. Ligand binding to the extracellular domain of receptor tyrosine kinases often provides an important regulatory step which determines the selectivity of intracellular signaling pathways. Combinatorial exon splicing can be used to shuffle, for example, intracellular domains of receptor molecules or signal transduction proteins, including SH2 domains, SH3 domains, kinase domains, phosphatase domains, and phospholipase domains. In another embodiment, variant of SH2 and SH3 domains are randomly shuffled with domains engineered as either protein kinase or phosphatase inhibitors and the combinatorial polypeptide library screened for the ability to block the function of, for example, the action of oncogenic proteins such as sic or ras.

Many techniques are known in the art for screening gene products of combinatorial libraries made by point mutations, and for screening cDNA libraries for gene products having a certain property. Such techniques will be generally applicable to screening the gene libraries generated by the present exon-shuffling methodology. The most widely used techniques for screening large gene libraries typically comprises cloning the gene library into replicable expression vectors, transforming appropriate cells with the resulting library of vectors, and expressing the combinatorial genes under conditions in which detection of a desired activity facilitates relatively easy isolation of the vector encoding the gene whose producted was detected. For instance, in the case of shuffling intracellular domains, phenstypic changes can be detected and used to isolate cells expressing a combinatoriallyderived gene product conferring the new phenotype. Likewise, interaction trap assays can be used in vivo to screen large polypeptide libraries for proteins able to bind a "bait" protein, or alternatively, to inhibit binding of two proteins.

For ribozymes, one illustrative embodiment comprises screening a ribozyme library for the ability of molecules to cleave an mRNA molecule and disrupt expression of a protein in such a manner as to confer some phenotypic change to the cell. Similarly, to assay the ability of novel autocatalytic introns to mediate splicing (e.g. see the group II domain shuffling described above) the ability of a combinatorial intron to mediate splicing between two exons can be detected by the ability to score for the protein product of the two exons when accurately spliced.

In yet another screening assay, the gene product, especially if its a polypeptide, is displayed on the surface of a cell or viral particle, and the ability of particular cells or viral particles to bind another molecule via this gene product is detected in a "panning assay". For example, the gene library can be cloned into the gene for a surface membrane protein of a bacterial cell, and the

resulting fusion protein detected on the surface of the bacteria (Ladner et al., WO 88/06630; Fuchs et al. (1991) Bio/Technology 9:1370-1371; and Goward et al. (1992) TIBS 18:136-140). In another embodiment, gene library is expressed as fusion protein on the surface of a viral particle. For instance, in the filamentous phage system, foreign peptide sequences can be expressed on the surface of infectious phage, thereby conferring two significant benefits. First, since these phage can be applied to affinity matrices at very high concentrations, large number of phage can be screened at one time. Second, since each infectious phage encodes the exon-shuffled gene product on its surface, if a particular phage is recovered from an affinity matrix in low yield, the phage can be amplified by another round of infection. The group of almost identical E.coli filamentous phages M13, fd, and fl are most often used in phage display libraries, as either of the phage glll or gVIII coat proteins can be used to generate fusion proteins without disrupting the ultimate packaging of the viral particle (Ladner et al. PCT publication WO 90/02909; Garrard et al., PCT publication WO 92/09690; Marks et al. (1992) J. Biol. Chem. 267:16007-16010; Griffths et al. (1993) EMBO J 12:725-734; Clackson et al. (1991) Nature 352:624-628; and Barbas et al. (1992) PNAS 89:4457-4461).

A. Antibody Repertoires

Mouse monoclonal antibodies are readily generated by the fusion of antibody-producing B lymphocytes with myeloma cells. However, for therapeutic applications, human monoclonal antibodies are preferred. Despite extensive efforts, including production of heterohybridomas, Epstein-Barr virus immortalization of human B cells, and "humanization" of mouse antibodies, no general method comparable to the Kohler-Milstein approach has emerged for the generation of human monoclonal antibodies.

Recently, however, techniques have been developed for the generation of antibody libraries in E. coli capable of expressing the antigen binding portions of immunoglobulin heavy and light chains. For example, recombinant antibodies have been generated in the form of fusion proteins containing membrane proteins such as peptidoglycan-assoicated lipoprotein (PAL), as well as fusion proteins with the capsular proteins of viral particles, or simply as secreted proteins which are able to cross the bacterial membrane after the addition of a bacterial leader sequence at their N-termini. (See, for example, Fuchs et al. (1991) Bio/Technology 9:1370-1372; Bettes et al. (1988) Science 240:1041-1043; Skerra et al. (1988) Science 240:1038-1041; Hay et al. (1992) Hum. Antibod. Hybridomas 3:81-85; and Barbas et al. International Publication No. W0 92/18019).

The display of antibody fragments on the surface of filamentous phage that encode the antibody gene, and the selection of phage binding to a particular antigen, offer a powerful means of generating specific antibodies in vitro. Typically, phage antibodies (phAbs) have been generated and expressed in bacteria by cloning repertoires of rearranged heavy and light chain V-genes into filamentous bacteriophage. Antibodies of a particular specificity can be selected from the phAb library by panning with antigen. The present intron-mediated combinatorial approach can be applied advantageously to the production of recombinant antibodies by providing antibody libraries not readily accessible by any prior technique. For instance, in contrast to merely sampling combinations of V.sub.H and V.sub.L chains, the present method allows the complementarity-determining regions (CDRs) and framework regions (FRs) themselves to be

randomly shuffled in order to create novel V.sub.H and V.sub.L regions which were not represented in the originally cloned rearranged V-genes.

Antibody variable domains consist of a .beta.-sheet framework with three loops of hypervariable sequences (e.g. the CDRs) (see FIG. 20A), and the antigen binding site is shaped by loops from both heavy (V.sub.H) and light (V.sub.L) domains. The loops create antigen binding sites of a variety of shapes, ranging from flat surfaces to pockets. For human V.sub.H domains, the sequence diversity of the first two CDRs are encoded by a repertoire of about 50 germline V.sub.H segments (Tomlinson et al. (1992) J. Mol. Biol. 227:). The third CDR is generated from the combination of these segments with about 30 D and six J segments (Ichihara et al. (1988) EMBO J 7: 4141-4150). The lengths of the first two CDRs are restricted, with the length being 6 amino acid residues for CDR1, 17 residues, and for CDR2. However, the length of CDR3 can differ significantly, with lengths ranging from 4 to 25 residues.

For human light chain variable domains, the sequence diversity of the first two CDRs and part of CDR3 are encoded by a repertoire of about 50 human V.sub..kappa. segments (Meindl et al. (1990) Eur. J Immunol. 20: 1855-1863) and >10 V.sub..lambda. segments (Chuchana et al. (1990) Eur. J. Immunol. 20: 1317-1325; and Combriato et al. (1991) Eur. J Immunol. 21: 1513-1522). The lengths of the CDRs are as follows, CDR1=11-14 residues; CDR2=8 residues; and CDR3 ranges from 6 to 10 residues for V.sub..kappa. genes and 9 to 13 for V.sub..lambda. genes.

The present invention contemplates combinatorial methods for generating diverse antibody libraries, as well as reagents and kits for carrying out such methods. In one embodiment, the present combinatorial approach can be used to recombine both the framework regions and CDRs to generate a library of novel heavy and light chains. In another embodiment, trans-splicing can be used to shuffle only the framework regions which flank specific CDR sequences. While both schemes can be used to generate antibodies directed to a certain antigen, the later strategy is particularly amenable to being used for "humanizing" non-human monoclonal antibodies.

The combinatorial units useful for generating diverse antibody repertoires by the present trans-splicing methods comprise exon constructs corresponding to fragments of various immunoglobulin variable regions flanked by intronic sequences that can drive their ligation. As illustrated in FIGS. 20B and 20C, the "exonic" sequences of the combinatorial units can be selected to encode essentially just a framework region or CDR; or can be generated to correspond to larger fragments which may include both CDR and FR sequences. The combinatorial units can be made by standard cloning techniques that manipulate DNA sequences into vectors which provide appropriate flanking intron fragments upon transcription. Alternatively, the combinatorial units can be generated using reverse-splicing, as described above, to specifically add intronic sequences to fragments of antibody transcripts.

Methods are generally known for directly obtaining the DNA sequence of the variable regions of any immunoglobulin chain by using a mixture of oligomer primers and PCR. For instance, mixed oligonucleotide primers corresponding to the 5' leader (signal peptide) sequences and/or FRI sequences and a conserved 3' constant region primer have been used for PCR amplification of the heavy and light chain variable regions from a number of human antibodies directed to, for

example, epitopes on HIV-I (gp 120, gp 42), digoxin, tetanus, immunoglobulins (rheumatoid factor), and MHC class I and II proteins (Larrick et al. (1991) Methods: Companion to Methods in Enzymology 2: 106-110). A similar strategy has also been used to amplify mouse heavy and light chain variable regions from murine antibodies, such as antibodies raised against human T cell antigens (CD3, CD6), carcino embryonic antigen, and fibrin (Larrick et al. (1991) Bio Techniques 11: 152-156).

In the present invention, RNA is isolated from mature B cells of, for example, peripheral blood cells, bone marrow, or spleen preparations, using standard protocols. First-strand cDNA is synthesized using primers specific for the constant region of the heavy chain(s) and each of the .kappa. and .lambda. light chains. Using variable region PCR primers, such as those shown in Table I below, the variable regions of both heavy and light chains are amplified (preferably in separate reactions) and ligated into appropriate expression vectors. The resulting libraries of vectors (e.g. one for each of the heavy and light chains) contain a variegated population of variable regions that can be transcribed to generate mRNA enriched for V.sub.H and V.sub.L transcripts. Using the reversal of splicing reaction, group I or group II introns can be used which are designed to insert immediately downstream of specific nucleotide sites corresponding to the last (carboxy terminal) 2-3 amino acid residues of each framework region. For example, as depicted in FIG. 20B, a set of group II Y-branched lariats can be utilized to specifically insert flanking group II intron fragments between each CDR sequence and the FR sequence immediately upstream. The exon binding sequence (EBS 1, and in some instances EBS2) of each Y-branched lariat is manipulated to create a panel of Y lariats based on sequence analysis of known framework regions (FR1-4). The intronic addition can be carried out simultaneously for all three FR/CDR boundaries, or at fewer than all three boundaries. For instance, the RNA transcripts can be incubated with Y lariats which drive insertion at only the FR1/CDR1 and FR2/CDR2 boundaries. The resulting intron-containing fragments can be reverse transcribed using a domain VI primer, and the cDNA amplified using PCR primers complementary to a portion of domain VI, a portion of domain I, and the leader sequence. Thus, the Leader,FR1(IVS 1-3) and (IVS 5,6)CDR1,FR2(IVS 1-3) constructs will be generated. Likewise, the RNA transcript can instead be incubated under reverse-splicing conditions with Y-branched lariats which are directed to insertion at the FR2/CDR2 and FR3/CDR3 boundaries, resulting in the (IVS 5,6)CDR2,FR3(IVS 1-3) and (IVS 5,6)CDR3,FR4 combinatorial units, which can then be isolated by reverse transcription and PCR using primers to sequences in domain I, domain VI, and the constant region.

TABLE I _____ Human Immunoglobulin Variable Region PCR Primers _____ 5' End Sense Human heavy chains Group A 5'-GGGAATTCATGGACTGGACCTGGAGG(AG)TC(CT)-- TCT(GT)C-3' (SEQ ID NO.11) Group B 5'-GGGAATTCATGGAG(CT)TTGGGCTGA(CG)CTGG(CG)-- TTTT-3' (SEQ ID NO:12) Group C 5'-GGGAATTCATG(AG)A(AC)(AT)ACT(GT)TG(GT)-- (AT)(CG)C(AT)(CT)(CG)CT(CT)CTG-3' (SEQ ID NO:13) Human .kappa. light chain 5'-GGGAATTCATGGACATG(AG)(AG)(AG)(AGT)(CT)CC-- (ACT)(ACG)G(CT)GT)CA(CG)CTT-3'(SEQ ID NO:14) Human .lambda. light chain 5'-GGGAATTCATG(AG)CCTG(CG)(AT)C(CT)CCTCTC(CT)-- T(CT)CT(CG)(AT)(CT)C-3' (SEQ ID NO:15) 3' End sense constant region Human IgM heavy chain 5'-CCAAGCTTAGACGAGGGGGAAAAGGGTT-3'(SEQ ID NO:16) Human IgG1 heavy chain

5'CCAAGCTTGGAGGAGGGTGCCAGGGGG-3'(SEQ ID NO:17) Human .lambda. light chain 5'-CCAAGCTTGAAGCTCCTCAGAGGAGGG-3'(SEQ ID NO:18) Human .kappa. light chain 5'-CCAAGCTTTCATCAGATGGCGGGAAGAT-3'(SEQ ID NO:19) _____ Murine Immunoglobulin Variable Region PCR Primers _____ 5' End Sense Leader (signal peptide) region (amino acids -20 to -13) Group A 5'-GGGGAATTCATG(GA)A(GC)TT(GC)(TG)GG(TC)T(AC)-- A(AG)CT(GT)G(GA)TT-3'(SEQ ID NO:20) Group B 5'-GGGGAATTCATG(GA)AATG(GC)A(GC)CTGGGT(CT)-- (TA)T(TC)CTCT-3'(SEQ ID NO:21) Framework 1 region (amino acids 1 to 8) 5'-GGGGAATTC(CG)AGGTG(CA)AGCTC(CG)(AT)(AG)(CG)-- A(AG)(CT)C(CG)GGG-3'(SEQ ID NO:22) 3' End sense constant region Mouse .gamma. constant region (amino acids 121 to 131) 5'-GGAAGCTTA(TC)CTCCACACACAGG(AG)(AG)CCAGTG-- GATAGAC-3'(SEQ ID NO:23) Mouse .kappa. light chain (amino acids 116 to 122) 5'-GGAAGCTTACTGGATGGTGGGAAGATGGA-3'(SEQ ID NO:24) _____ Bases in parentheses represent substitutions at a given residue. EcoRI an HindIII sites are underlined.

The Leader, FR1 (IVS 1-3) transcripts can be linked to an insoluble resin by standard techniques, and each set of combinatorial units (CDR1/FR2, CDR2/FR3, CDR3/FR4) can be sequentially added to the resin-bound nucleic acid by incubation under trans-splicing conditions, with unbound reactants washed away between each round of addition. After addition of the (IVS 5,6)CDR3,FR4 units to the resin bound molecules, the resulting trans-spliced molecule can be released from the resin, reverse-transcribed and PCR amplified using primers for the leader sequence and constant region, and subsequently cloned into an appropriate vector for generating a screenable population of antibody molecules.

Taking the dissection of the variable regions one step further, a set of exon libraries can be generated for ordered combinatorial ligation much the same as above, except that each combinatorial unit is flanked at its 5' end with an intron fragment that is unable to drive a trans-splicing reaction with the intron fragment at its 3' end. As described above (section II) with regard to ordered gene assembly, each combinatorial unit is effectively protected from addition by another unit having identical flanking intron fragments. The 5' and 3' flanking intronic sequences can be of the same group, but from divergent enough classes (i.e. group IIA versus group IIB) or divided in such a way that intermolecular complementation and assembly of an active splicing complex cannot occur; or the intron fragments can simply be from different groups (e.g. group I versus group II).

As illustrated in FIG. 20C, the combinatorial units of FIG. 20B can be generated with Y lariats derived from group IIA intron fragments (hence the designation "IVS-A-5, 6"). Each CDR is then split from the downstream framework region using a Y-branched lariat derived from a group IIB intron having a divergent enough domain V that neither combination of (IVS-A-5,6) and (IVS-B-1-3) or (IVS-B-5,6) and (IVS-A-1-3) results in a functional splicing complex. In order to avoid the need to determine the sequence of each of the cloned CDRs, the exon-binding sites of the IIB intron lariats can be constructed to match the much less variable nucleotide sequences corresponding to the first (amino terminal) 2-3 a.a. residues of each of the framework regions (FR2-4). The resulting constructs include internal exon units of the general formula

(IVS-A-5,6) CDR (IVS-B-1-3) and (IVS-B-5,6) FR (IVS-A-1-3), with each CDR containing an extra 2-3 a.a. residues from the FR which previously flanked it. Thus, by sequentially adding each pool of combinatorial units to the resin-immobilized FR1, an ordered combinatorial ligation of variegated populations of CDRs and FRs can be carried out to produce a library of variable region genes in which both the CDRs and FRs have been independently randomized.

Furthermore, CDR combinatorial units can be generated which are completely random in sequence, rather than cloned from any antibody source. For instance, a plasmid similar to pINV1 (described herein) can be used to create a set of random CDR sequences of a given length and which are flanked by appropriate intronic fragments. In an illustrative embodiment, the plasmid includes restriction endonuclease sites in each of the 5' and 3' flanking intron sequences such that oligonucleotides having the CDR coding sequence can be cloned into the plasmid. For example, a degenerate oligonucleotide can be synthesized for CDR1 which encodes all possible amino acid combinations for the 6 a.a. sequence. The nucleotide sequences which flank the CDR-encoding portion of the oligonucleotide comprise the flanking intron sequences necessary to allow ligation of the degenerate oligonucleotide into the plasmid and reconstitute a construct which would produce a spliceable transcript. To avoid creation of stop codons which can result when codons are randomly synthesized using nucleotide monomers, "dirty bottle" synthesis can instead be carried out using a set of nucleotide trimers which encode all 20 amino acids.

With slight modification, the present ordered combinatorial ligation can be used to efficiently humanize monoclonal antibodies of non-human origin. The CDRs from the monoclonal antibody can be recombined with human framework region libraries (e.g. an FR1 library, an FR2 library, etc.) to produce a combinatorial population of variable regions in which the CDR sequences are held constant, but each of the framework regions have been randomized. The variable regions can be subsequently fused with sequences corresponding to the appropriate human constant regions, and the antibodies resulting from heavy and light chain association can be screened for antigen binding using standard panning assays such as phage display. In contrast to contemporary humanization schemes which require the practitioner to prejudically choose a particular human scaffold into which the CDRs are grafted, the present technique provides a greater flexibility in choosing appropriate human framework regions which do not adversely affect antigen binding by the resultant chimeric antibody.

To illustrate, the variable regions of both the heavy and light chains of a mouse monoclonal antibody can be cloned using primers as described above. The sequence of each CDR can be obtained by standard techniques. The CDRs can be cloned into vectors which provide appropriate flanking intronic sequences, or alternatively, isolated by reverse-splicing with Y-branched lariats designed to insert precisely at each FR/CDR and CDR/FR boundary. As described above, the particular intronic fragments provided with each murine CDR and each human FR construct can be selected to disfavor multiple ligations at each step of addition to a resin bound nucleic acid. The library of human heavy chain leader, FR1 (IVS-A1-3) constructs can be immobilized on a resin, and in a first round of ligation, the heavy chain murine (IVS-A-5,6) CDR1 (IVS-B-1-3) construct is added under trans-splicing conditions. Un-ligated combinatorial units are washed away, and the library of human heavy chain (IVS-B-5,6) FR2 (IVS-A-1-3) units are admixed and trans-spliced to the resin-bound nucleic acids terminating with the murine CDR construct. This process is carried out for the remaining murine CDR and

human FR units of the heavy chain, and a similar process is used to construct combinatorial light chain chimeras as well. The resulting chimeric heavy and light chains can be cloned into a phage display library, and the phabs screened in a panning assay to isolate humanized antibodies (and their genes) which bind the antigen of interest.

B. Combinatorial Enzyme Libraries

Plasminogen activators (PAs) are a class of serine proteases that convert the proenzyme plasminogen into plasmin, which then degrades the fibrin network of blood clots. The plasminogen activators have been classified into two immunologically unrelated groups, the urokinase-type PAs (u-PA) and the tissue-type PA (tPA), with the later activator being the physiological vascular activator. These proteins, as well as other proteases of the fibrinolytic pathway, are composed of multiple structural domains which appear to have evolved by genetic assembly of individual subunits with specific structural and/or functional properties. For instance, the amino terminal region of tPA(SEQ ID No: 25) is composed of multiple structural/functional domains found in other plasma proteins, including a "finger-like domain" homologous to the finger domains of fibronectin, an "epidermal growth factor domain" homologous to human EGF, and two disulfide-bonded triple loop structures, commonly referred to as "kringle domains", homologous to the kringle regions in plasminogen. The region comprising residues 276-527 (the "catalytic domain" is homologous to that of other serine proteases and contains the catalytic triad. In addition, the gene for tPA encodes a signal secretion peptide which directs secretion of the protein into the extracellular environment, as well as a pro-sequence which is cleaved from the inactive form of the protease (the "plasminogen") to active tPA during the fibrinolytic cascade.

These distinct domains in tPA are involved in several functions of the enzyme, including its binding to fibrin, stimulation of plasminogen activation by fibrin, and rapid in vivo clearance. Approaches used to characterize the functional contribution of these structural domains include isolation of independent structural domains as well as the production of variant proteins which lack one or more domains. For example, the fibrin selectivity of tPA is found to be mediated by its affinity for fibrin conferred by the finger-like domain and by at least one of the kringle domains.

The present combinatorial method can be used to generate novel plasminogen activators having superior thrombolytic properties, by generating a library of proteins by RNA-splicing mediated shuffling of the domains of plasma proteins. As described below, one mode of generating the combinatorial library comprises the random trans-splicing of a mixture of exons corresponding to each of the domains of the mature tPA protein. Briefly, a cDNA clone of tPA was obtained and, through the use of specific PCR amplimers, each of the 5 protein domains was amplified and isolated. Each of these amplified domains was then separately cloned into a plasmid as an exon module such that the 5' end of the exon is preceded by group II domains V-VI, and the 3' end of the exon is followed by group II domains I-III. In addition, the IBS 1 site of each of the exon was mutated in order to facilitate base pairing with the EBS 1 sequence of the 3' flanking intron fragment. Transcription of the resulting construct thus produces RNA transcripts of the general formula (IVS 5,6)-Exon-(IVS 1-3). Mixture of these transcripts under trans-splicing conditions can result in random ligation of the exons to one and other and assembly of the

combinatorial gene library which can subsequently be screened for fibrinolytic activity.

Moreover, combinatorial units can be generated from other proteins, including proteins having no catalytic role in blood clotting or fibrinolysis. For example, a library of catalytic domains can be generated from other thrombolytic proteases, blood clotting factors, and other proteases having peptidic activity similar to the typsin-like activity of tPA(SEQ ID NO:25) Likewise, libraries of splicing constructs can be derived from EGF-like domains, finger-like domains, kringle domains, and Calcium-binding domains from a vast array of proteins which contain such moieties.

EXAMPLE 4

Construction of plasmid TPA-KS.sup.+

The cDNA clone of the human tissue plasminogen activator (tPA) gene (pETPFR) was obtained from the ATCC collection (ATCC 40403; and U.S. Pat. No. 4,766,075). The entire cDNA clone was amplified by PCR using primers 5'-ACGATGCATGCTGGAGA GAAAACCTCTGCG(SEQ ID NO:26) and 5'-ACGATGCATTCTGTAGAGAAGCACTGCGCC(SEQ ID No: 27). TPA sequences from 70 base pairs (bp) upstream of the translation initiation site (AUG) to 88 bp downstream of the translation termination site (TGA) were amplified (SEQ. ID No. 3). In addition, the primers added Nsi I sites to both ends of the amplified DNA. The amplified DNA was cut with Nsi I and ligated into the KS.sup.+ vector that had been cut with Pst I. A clone TPA-KS.sup.+, was isolated with the insert oriented such that in vitro transcription with T.sub.7 RNA polymerase yields an RNA that is the same polarity as the tPA mRNA.

EXAMPLE 5

Construction of plasmid INV-KX

Two unique restriction sites were added to the pINV1 plasmid (SEQ ID NO. 1) by site directed mutagenesis, to facilitate insertion of portions of the tPA clone. A Kpn I site (GGTACC) was inserted at precisely the boundary between the end of the intron and the beginning of E3. An Xho I site was added to E5 by changing the sequence GTGGGA to a Xho I site (CTCGAG). Thus, the last seven bp of the exon were unchanged, but the six preceding base pairs were changed to create a Xho I site. The resulting plasmid is termed here INV-KX.

EXAMPLE 6

Construction of plasmid INV-K(K1)X

The region of the TPA cDNA clone that encodes the kringle-1(K1) domain was amplified by PCR. The primers added a Kpn I site at the upstream end of the domain and a Xho I site to the downstream end. The amplified DNA was cut with Kpn I and Xho I and ligated into INV-KX such that the K1 sequences replaced the E3,E5 exon sequences.

EXAMPLE 7

Construction of plasmid (IVS 5,6)K1(IVS1-3)

Oligonucleotide splints were used in a site-directed mutagenesis experiment to change the sequences at the boundaries of the INV-KX derived introns and the K1 exon as well as to remove the Kpn I and Xho I sites. The sequences were changed such that the intron sequences of domain 6 are directly followed by kringle domain sequences ACC AGG GCC and kringle sequences TCT GAG GGA precede the intron sequences of domain 1. In addition, the sequence of the EBS 1 sequence in domain 1 was changes to TCCCTCA (this sequence is homologous to the last 7 nt of K1 (TGAGGGA). Thus, the resulting transcript, (IVS5,6)K1(IVS1-3), contains complementary IBS1 and EBS1 sequences.

As an alternate construct, an oligonucleotide splint was used to remove the extra nucleotide sequences, e.g. the Kpn I site, between domain 6 and the 5' end of the kringle domain. However, the oligonucleotide primer used to change the EBS 1 sequence in domain 1 to TCCCTCA did not remove the Xho I site, leaving an extra 13 nucleotides (CTCGAGCATTTTC(SEQ ID No: 28) between the 3' end of the kringle domain and domain I of the flnaking intron fragment. It is not believed that this additional stretch of nucleotides will have any significant effect on splicing (see, for example, Jacquier et al. (1991) J Mol Biol 219:415-428).

EXAMPLE 8

Construction ofplasmid GrII-Sig

Two oligonucleotide primers were used to change the IBS 1 sequence of pINV1 to TGTCAAA and the EBS 1 sequence to TTTGACA. Thus, the last seven nucleotides of E5 were changes to the sequence of the last 7 nucleotides of TPA fibronectin finger like domain and the EBS 1 sequence was made complementary. The resulting plasmid is termed here GrII-Sig.

EXAMPLE 9

Construction of plasmid SIG(IVS1-3)

The plasmid SIG(IVS1-3) contains the first two protein domains of TPA (the signal sequence and the finger domain) followed by group II intron domains 1-3. It was made by the reversal of splicing. Plasmid GrII-Sig (Example 8) was linearized with Hind III and RNA made using T.sub.7 polymerase in vitro. The RNA was incubated under self splicing conditions for two hours and the products fractionated on an acrylamide gel. The Sig(Y) molecule (a Y-branched lariat intron comprising domains 5 and 6 joined to domains 1 through 3 by a 2'-5' phosphodiester bond) was gel purified. This molecule was the "enzyme" used for the reverse-splicing reaction. The substrate was made by cutting TPA-KS.sup.+ DNA (Example 4) with Sty I, which cuts 17 bp downstream of the end of the finger domain. A 404 nt RNA was made using T.sub.7 polymerase. The enzyme and substrate were mixed and incubated under splicing conditions for two hours. By the reversal of splicing, the Sig(Y) RNA attacked the substrate to yield the signal plus finger region followed by intron domains 1 through 3. A cDNA copy of the molecule was

made using reverse transcriptase and amplified by PCR. It was cloned into the PBS vector in the T.sub.7 orientation. The clones analyzed each showed precise joining of the coding sequence to the group II intron sequence. Thus, the nucleotide sequence of the EBS1 was sufficient to direct exact integration of the intronic IVS(1-3) fragment.

EXAMPLE 10

Construction of other shuffling clones

Clones with each of the other three protein domains (growth factor (GF) domain, kringle 2 (K2) domain and catalytic (cat) domain), flanked by group II intron sequences, can also be made by either standard cloning methods or by the reversal of splicing method, as described above, to yield constructs corresponding to (IVS5,6)FG(IVS 1-3), (IVS5,6)K2(IVS 1-3), and (IVS5,6)cat or (IVS5,6)cat(IVS1-3).

To further illustrate, the plasmid (IVS5,6)cat was generated by reversal of splicing as in Example 9. Briefly, the Y-branched intron of the pY7 construct (see Example 3) was generated by cutting the pY7 plasmid with HindIll, producing RNA with T.sub.7, and incubating the RNA under self-splicing conditions for 1.5 hours. The products were fractionated on an acylamide gel. The Y7 molecule (a Y-branched intron) was gel purified. This molecule was used for the reverse-splicing reaction. The substrate for this reaction was generated by cutting a plasmid containing the tPA catalytic domain with HindIII and transcribing the linear plasmid with T.sub.7. The Y-branched enzyme and tPA substrate RNAs were mixed and incubated under reverse-splicing conditions for 4 hours. By the reversal of splicing, the Y7 RNA attacked the substrate at a site (IBS1) just upstream of the coding sequence for the catalytic domain to yield intron domains 5 and 6 followed by the tPA protease domain. A cDNA copy of the molecule was made using reverse transcriptase and amplified by PCR. It was cloned into a PBS vector in the T.sub.7 orientation. Two independent clones were characterized by DNA sequence analysis. Both clones had the group II intron sequences precisely joined to the tPA sequences. Thus, the nucleotide sequence of the fusion protin was 5'ATCGGGAT/ACCTGCGG a(SEQ ID No: 29) (intron/exon, respectively).

EXAMPLE 11

Generation of library

RNA transcripts are made for each of the tPA combinatorial units, SIG(IVS1-3), (IVS5,6)K1(IVS 1-3), (IVS5,6)K2(IVS 1-3), (IVS5,6)GF(IVS 1-3), and (IVS5,6)cat(IVS 1-3). The transcripts are mixed and incubated under trans-splicing conditions. The resulting combinatorial RNA molecules can be reverse-transcribed to cDNA using primers complementary to sequences in the intron domains I-III, and the cDNA amplified by PCR using a similar primer and a primer to the tPA signal sequence. The amplified cDNAs can subsequently be cloned into suitable expressions vectors to generate an expressions library, and the library screened for fibrinolytic activity by standard assays.

All of the above-cited references and publications are hereby incorporated by reference.

Equivalents

Those skilled in the art will recognize, or be able to ascertain using no more than routine experimentation, numerous equivalents to the specific methods and reagents described herein. Such equivalents are considered to be within the scope of this invention and are covered by the following claims.

SEQUENCE LISTING (1) GENERAL INFORMATION: (iii) NUMBER OF SEQUENCES: 29 (2) INFORMATION FOR SEQ ID NO:1: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 4539 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: both (ii) MOLECULE TYPE: other nucleic acid (ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 969..1259 (D) OTHER INFORMATION: /product="E3 exon" (ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 1290..1559 (D) OTHER INFORMATION: /product="E5 exon" (xi) SEQUENCE DESCRIPTION: SEQ ID NO:1:

TCGCGCGTTTCGGTGATGACGGTGAAAACCTCTGACACATGCAGCTCCCGGAGACG
GTCA60

CAGCTTGTCTGTAAGCGGATGCCGGGAGCAGACAAGCCCGTCAGGGCGCGTCAGCG
GGTG120

TTGGCGGGTGTCGGGGCTGGCTTAACTATGCGGCATCAGAGCAGATTGTACTGAGA
GTGC180

ACCATATGCGGTGTGAAATACCGCACAGATGCGTAAGGAGAAAATACCGCATCAGG
CGAC240

GCGCCCTGTAGCGGCGCATTAAGCGCGGCGGGTGTGGTGGTTACGCGCAGCGTGAC
CGCT300

ACACTTGCCAGCGCCCTAGCGCCCGCTCCTTTCGCTTTCTTCCCTTCCTTTCTCGCCA
CG360

TTCGCCGGCTTTCCCCGTCAAGCTCTAAATCGGGGGCTCCCTTTAGGGTTCCGATTTA
GT420

GCTTTACGGCACCTCGACCCCAAAAAACTTGATTAGGGTGATGGTTCACGTAGTGGG
CCA480

TCGCCCTGATAGACGGTTTTTCGCCCTTTGACGTTGGAGTCCACGTTCTTTAATAGTG
GA540

CTCTTGTTCCAAACTGGAACAACACTCAACCCTATCTCGGTCTATTCTTTTGATTTAT
AA600

GGGATTTTGCCGATTTCGGCCTATTGGTTAAAAAATGAGCTGATTTAACAAAAATTT
AAC660

GCGAATTTTAACAAAATATTAACGCTTTACAATTTCGCCATTCGCCATTCAGGCTGC
GCA720

ACTGTTGGGAAGGGCGATCGGTGCGGGCCTCTTCGCTATTACGCCAGCTGGCGAAA
GGGG780

GATGTGCTGCAAGGCGATTAAGTTGGGTAACGCCAGGGTTTTCCCAGTCACGACGTT
GTA840

AAACGACGGCCAGTGAATTGTAATACGACTCACTATAGGGCGAATTCGAGCTCGTG

AGCC900
GTATGCGATGAAAGTCGCACGTACGGTTCTTACCGGGGGAAAACTTGTAAAGGTCT
ACCT960
ATCGGGATACTATGTATTATCAATGGGTGCTATTTTCTCTTTATTTGCAGGATACTAC
TA1020
TTGAAGTCCTCAAATTTTAGGTTTAAACTATAATGAAAAATTAGCTCAAATTCAATT
CTG1080
ATTAATTTTCATTGGGGCTAATGTTATTTTCTTCCCAATGCATTTCTTAGGTATTAAT
GG1140
TATGCCTAGAAGAATTCCTGATTATCCTGATGCTTTCGCAGGATGAAATTATGTCGC
TTC1200
TATTGGTTCATTCATTGCACTATTATCATTATTCTTATTTATCTATATTTTATATGATC
C1260
TCTAGAGTCGACCTGCAGCCCAAGCTGGGGATCACATCATATGTATATTGTAGGATT
AGA1320
TGCAGATACTAGAGCATATTTCCTATCCGCACTGATGATTATTGCAATTCCAACAGG
AAT1380
TAAAATCTTTTCTTGATTAGCCCTGATCTACGGTGGTTCAATTAGATTAGCACTACCT
AT1440
GTTATATGCAATTGCATTCTTATTCTTATTCACAATGGGTGGTTTAACTGGTGTTGCC
TT1500
AGCTAACGCCTCATTAGATGTGGCATTCCACGATACTTACTACGTGGTGGGACATTT
TCG1560
AGCGGTCTGAAAGTTATCATAAATAATATTTACCATATAATAATGGATAAATTATAT
TTT1620
TATCAATATAAGTCTAATTACAAGTGTATTAAAATGGTAACATAAATATGCTAAGCT
GTA1680
ATGACAAAAGTATCCATATTCTTGACAGTTATTTTATATTATAAAAAAAAGATGAAG
GAA1740
CTTTGACTGATCTAATATGCTCAACGAAAGTGAATCAAATGTTATAAAATTACTTAC
ACC1800
ACTAATTGAAAACCTGTCTGATATTCAATTATTATTTATTATTATATAATTATATAAT
AA1860
TAAATAAAATGGTTGATGTTATGTATTGGAAATGAGCATACGATAAATCATATAACC
ATT1920
AGTAATATAATTTGAGAGCTAAGTTAGATATTTACGTATTTATGATAAAACAGAATA
AAC1980
CCTATAAATTATTATTATTAATAATAAAAAATAATAATAATACCAATATATATATTA
TTT2040
AATTTATTATTATTATATTAATAAAATTTAATATATATTATAAATAATTATTGGATTA
AG2100
AAATATAATATTTTATAGAAATTTTCTTTATATTTAGAGGGTAAAAGATTGTATAAA
AAG2160
CTAATGCCATATTGTAATGATATGGATAAGAATTATTATTCTAAAGATGAAAATCTG
CTA2220
ACTTATACTATAGGGGGGATCCTCTAGAGTCGACCTGCAGGCATGCAAGCTTTTGTT

CCC2280
TTTAGTGAGGGTTAATTTCGAGCTTGGCGTAATCATGGTCATAGCTGTTTCCTGTGTG
AA2340
ATTGTTATCCGCTCACAATTCCACACAACATACGAGCCGGAAGCATAAAGTGTAAA
GCCT2400
GGGGTGCCTAATGAGTGAGCTAACTCACATTAATTGCGTTGCGCTCACTGCCCGCTT
TCC2460
AGTCGGGAAACCTGTCGTGCCAGCTGCATTAATGAATCGGCCAACGCGCGGGGAGA
GGCG2520
GTTTGCGTATTGGGCGCTCTTCCGCTTCCTCGCTCACTGACTCGCTGCGCTCGGTCGT
TC2580
GGCTGCGGCGAGCGGTATCAGCTCACTCAAAGGCGGTAATACGGTTATCCACAGAA
TCAG2640
GGGATAACGCAGGAAAGAACATGTGAGCAAAAGGCCAGCAAAAGGCCAGGAACCG
TAAAA2700
AGGCCGCGTTGCTGGCGTTTTTCCATAGGCTCCGCCCCCCTGACGAGCATCACAAAA
ATC2760
GACGCTCAAGTCAGAGGTGGCGAAACCCGACAGGACTATAAAGATACCAGGCGTTT
CCCC2820
CTGGAAGCTCCCTCGTGCGCTCTCCTGTTCCGACCCTGCCGCTTACCGGATACCTGTC
CG2880
CCTTTCTCCCTTCGGGAAGCGTGGCGCTTTCTCATAGCTCACGCTGTAGGTATCTCAG
TT2940
CGGTGTAGGTCGTTCGCTCCAAGCTGGGCTGTGTGCACGAACCCCCCGTTCAGCCCG
ACC3000
GCTGCGCCTTATCCGGTAACTATCGTCTTGAGTCCAACCCGGTAAGACACGACTTAT
CGC3060
CACTGGCAGCAGCCACTGGTAACAGGATTAGCAGAGCGAGGTATGTAGGCGGTGCT
ACAG3120
AGTTCTTGAAGTGGTGGCCTAACTACGGCTACACTAGAAGGACAGTATTTGGTATCT
GCG3180
CTCTGCTGAAGCCAGTTACCTTCGGAAAAAGAGTTGGTAGCTCTTGATCCGGCAAAC
AAA3240
CCACCGCTGGTAGCGGTGGTTTTTTTGTTTGCAAGCAGCAGATTACGCGCAGAAAAA
AAG3300
GATCTCAAGAAGATCCTTTGATCTTTTCTACGGGGTCTGACGCTCAGTGGAACGAAA
ACT3360
CACGTTAAGGGATTTTGGTCATGAGATTATCAAAAAGGATCTTCACCTAGATCCTTT
TAA3420
ATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTGGTCTGACA
GTT3480
ACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCTATTTCGTTCATCCAT
AG3540
TTGCCTGACTCCCCGTCGTGTAGATAACTACGATACGGGAGGGCTTACCATCTGGCC
CCA3600
GTGCTGCAATGATACCGCGAGACCCACGCTCACCGGCTCCAGATTTATCAGCAATAA

ACC3660
AGCCAGCCGGAAGGGCCGAGCGCAGAAGTGGTCCTGCAACTTTATCCGCCTCCATC
CAGT3720
CTATTAATTGTTGCCGGGAAGCTAGAGTAAGTAGTTCGCCAGTTAATAGTTTGCGCA
ACG3780
TTGTTGCCATTGCTACAGGCATCGTGGTGTCACGCTCGTCGTTTGGTATGGCTTCATT
CA3840
GCTCCGGTTCCCAACGATCAAGGCGAGTTACATGATCCCCCATGTTGTGCAAAAAAG
CGG3900
TTAGCTCCTTCGGTCCTCCGATCGTTGTCAGAAGTAAGTTGGCCGCAGTGTTATCACT
CA3960
TGGTTATGGCAGCACTGCATAATTCTCTTACTGTCATGCCATCCGTAAGATGCTTTTC
TG4020
TGACTGGTGAGTACTCAACCAAGTCATTCTGAGAATAGTGTATGCGGCGACCGAGTT
GCT4080
CTTGCCCGGCGTCAATACGGGATAATACCGCGCCACATAGCAGAACTTTAAAAGTG
CTCA4140
TCATTGGAAAACGTTCTTCGGGGCGAAAACTCTCAAGGATCTTACCGCTGTTGAGAT
CCA4200
GTTCGATGTAACCCACTCGTGCACCCAACTGATCTTCAGCATCTTTTACTTTCACCAG
CG4260
TTTCTGGGTGAGCAAAAACAGGAAGGCAAAATGCCGCAAAAAAGGGAATAAGGGC
GACAC4320
GGAAATGTTGAATACTCATACTCTTCCTTTTTCAATATTATTGAAGCATTTATCAGGG
TT4380
ATTGTCTCATGAGCGGATACATATTTGAATGTATTTAGAAAAATAAACAAATAGGGG
TTC4440
CGCGCACATTTCCCCGAAAAGTGCCACCTGACGTCTAAGAAACCATTATTATCATGA
CAT4500 TAACCTATAAAAATAGGCGTATCACGAGGCCCTTTCGTC4539 (2)
INFORMATION FOR SEQ ID NO:2: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH:
2939 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: both
(ii) MOLECULE TYPE: other nucleic acid (ix) FEATURE: (A) NAME/KEY: misc.sub.--
feature (B) LOCATION: 2448..2657 (D) OTHER INFORMATION: /product="b-globin exon 2"
(ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 2667..2814 (D) OTHER
INFORMATION: /product="b-globin exon 1" (ix) FEATURE: (A) NAME/KEY: misc.sub.--
feature (B) LOCATION: 2815..2890 (D) OTHER INFORMATION: /product="intron sequence"
(ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 2390..2447 (D) OTHER
INFORMATION: /product="intron sequence" (xi) SEQUENCE DESCRIPTION: SEQ ID NO:2:
TATAGTGTCACCTAAATCGTATGTGTATGATACATAAGGTTATGTATTAATTGTAGC
CGC60
GTTCTAACGACAATATGTCCATATGGTGCACTCTCAGTACAATCTGCTCTGATGCCG
CAT120
AGTTAAGCCAGCCCCGACACCCGCCAACACCCGCTGACGCGCCCTGACGGGCTTGT
CTGC180
TCCCGGCATCCGCTTACAGACAAGCTGTGACCGTCTCCGGGAGCTGCATGTGTCAGA
GGT240

TTTCACCGTCATCACCGAAACGCGCGAGACGAAAGGGCCTCGTGATACGCCTATTTT
TAT300
AGGTTAATGTCATGATAATAATGGTTTCTTAGACGTCAGGTGGCACTTTTCGGGGAA
ATG360
TGCGCGGAACCCCTATTTGTTTATTTTTCTAAATACATTCAAATATGTATCCAGAGTA
TG420
AGTATTCAACATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGAGAGTATGAGTATTCA
AC480
ATTTCCGTGTCGCCCTTATTCCCTTTTTTGCGGCATTTTGCCTTCCTGTTTTTGCTCAC
C540
CAGAAACGCTGGTGAAAGTAAAAGATGCTGAAGATCAGTTGGGTGCACGAGTGGGT
TACA600
TCGAACTGGATCTCAACAGCGGTAAGATCCTTGAGAGTTTTCGCCCCGAAGAACGTT
TTC660
CAATGATGAGCACTTTTAAAGTTCTGCTATGTGGCGCGGTATTATCCCGTATTGACG
CCG720
GGCAAGAGCAACTCGGTCGCCGCATACACTATTCTCAGAATGACTTGGTTGAGTACT
CAC780
CAGTCACAGAAAAGCATCTTACGGATGGCATGACAGTAAGAGAATTATGCAGTGCT
GCCA840
TAACCATGAGTGATAACACTGCGGCCAACTTACTTCTGACAACGATCGGAGGACCG
AAGG900
AGCTAACCGCTTTTTTGCACAACATGGGGGATCATGTAACTCGCCTTGATCGTTGGG
AAC960
CGGAGCTGAATGAAGCCATACCAAACGACGAGCGTGACACCACGATGCCTGTAGCA
ATGG1020
CAACAACGTTGCGCAAACTATTAACTGGCGAACTACTTACTCTAGCTTCCCGGCAAC
AAT1080
TAATAGACTGGATGGAGGCGGATAAAGTTGCAGGACCACTTCTGCGCTCGGCCCTTC
CGG1140
CTGGCTGGTTTATTGCTGATAAATCTGGAGCCGGTGAGCGTGGGTCTCGCGGTATCA
TTG1200
CAGCACTGGGGCCAGATGGTAAGCCCTCCCGTATCGTAGTTATCTACACGACGGGG
AGTC1260
AGGCAACTATGGATGAACGAAATAGACAGATCGCTGAGATAGGTGCCTCACTGATT
AAGC1320
ATTGGTAACTGTCAGACCAAGTTTACTCATATATACTTTAGATTGATTTAAAACTTCA
TT1380
TTTAATTTAAAAGGATCTAGGTGAAGATCCTTTTTGATAATCTCATGACCAAAATCC
CTT1440
AACGTGAGTTTTCGTTCCACTGAGCGTCAGACCCCGTAGAAAAGATCAAAGGATCTT
CTT1500
GAGATCCTTTTTTTCTGCGCGTAATCTGCTGCTTGCAAACAAAAAAACCACCGCTAC
CAG1560
CGGTGGTTTGTTTGCCGGATCAAGAGCTACCAACTCTTTTTCCGAAGGTAACTGGCT
TCA1620

GCAGAGCGCAGATACCAAATACTGTCCTTCTAGTGTAGCCGTAGTTAGGCCACCACT
TCA1680
AGAACTCTGTAGCACCGCCTACATACCTCGCTCTGCTAATCCTGTTACCAGTGGCTG
CTG1740
CCAGTGGCGATAAGTCGTGTCTTACCGGGTTGGACTCAAGACGATAGTTACCGGATA
AGG1800
CGCAGCGGTCGGGCTGAACGGGGGGTTCGTGCACACAGCCCAGCTTGGAGCGAACG
ACCT1860
ACACCGAACTGAGATACCTACAGCGTGAGCATTGAGAAAGCGCCACGCTTCCCGAA
GGGA1920
GAAAGGCGGACAGGTATCCGGTAAGCGGCAGGGTCGGAACAGGAGAGCGCACGAG
GGAGC1980
TTCCAGGGGGAAACGCCTGGTATCTTTATAGTCCTGTCGGGTTTCGCCACCTCTGACT
TG2040
AGCGTCGATTTTTGTGATGCTCGTCAGGGGGGCGGAGCCTATGGAAAAACGCCAGC
AACG2100
CGGCCTTTTTACGGTTCCTGGCCTTTTGCTGGCCTTTTGCTCACATGTTCTTTCCTGCG
T2160
TATCCCCTGATTCTGTGGATAACCGTATTACCGCCTTTGAGTGAGCTGATACCGCTCG
CC2220
GCAGCCGAACGACCGAGCGCAGCGAGTCAGTGAGCGAGGAAGCGGAAGAGCGCCC
AATAC2280
GCAAACCGCCTCTCCCCGCGCGTTGGCCGATTCATTAATGCAGGTTAACCTGGCTTA
TCG2340
AAATTAATACGACTCACTATAGGGAGACCGGCCTCGAGCAGCTGAAGCTTTGGGTTT
CTG2400
ATAGGCACTGACTCTCTCTGCCTATTGGTCTATTTTCCCACCCTTAGGCTGCTGGTGG
TC2460
TACCCTTGGACCCAGAGGTTCTTTGAGTCCTTTGGGGATCTGTCCACTCCTGATGCTG
TT2520
ATGGGCAACCCTAAGGTGAAGGCTCATGGCAAGAAAGTGCTCGGTGCCTTTAGTGA
TGGC2580
CTGGCTCACCTGGACAACCTCAAGGGCACCTTTGCCACACTGAGTGAGCTGCACTGT
GAC2640
AAGCTGCACGTGGATCCCCCTGAAGCTTGCTTACATTTGCTTCTGACACAACTGTGTT
CA2700
CTAGCAACCTCAAACAGACACCATGGTGCACCTGACTCCTGAGGAGAAGTCTGCCG
TTAC2760
TGCCCTGTGGGGCAAGGTGAACGTGGATGAAGTTGGTGGTGAGGCCCTGGGCAGGT
TGGT2820
ATCAAGGTTACAAGACAGGTTTAAGGAGACCAATAGAAACTGGGCATGTGGAGACA
GAGA2880
AGACTCTTGGGATCCCCGGGTACCGAGCTCGAATTCATCGATGATATCAGATCTGGT
TC2939 (2) INFORMATION FOR SEQ ID NO:3: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 2162 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: double (D) TOPOLOGY: both (ii) MOLECULE TYPE: other nucleic acid (ix) FEATURE: (A)

NAME/KEY: misc.sub.-- feature (B) LOCATION: 82..334 (D) OTHER INFORMATION: /product="Signal Sequence and Finger-like domain" (ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 335..447 (D) OTHER INFORMATION: /product="EGF-like domain" (ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 448..714 (D) OTHER INFORMATION: /product="Kringle-1 domain" (ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 715..972 (D) OTHER INFORMATION: /product="Kringle-2 domain" (ix) FEATURE: (A) NAME/KEY: misc.sub.-- feature (B) LOCATION: 973..2162 (D) OTHER INFORMATION: /product="Catalytic domain" (xi) SEQUENCE DESCRIPTION: SEQ ID NO:3:

TGAGCACAGGGCTGGAGAGAAAACCTCTGCGAGGAAAGGGAAGGAGCAAGCCGTG
AATTT60
AAGGGACGCTGTGAAGCAATCATGGATGCAATGAAGAGAGGGCTCTGCTGTGTGCT
GCTG120
CTGTGTGGAGCAGTCTTCGTTTCGCCCAGCCAGGAAATCCATGCCCGATTCAGAAGA
GGA180
GCCAGATCTTACCAAGTGATCTGCAGAGATGAAAAAACGCAGATGATATACCAGCA
ACAT240
CAGTCATGGCTGCGCCCTGTGCTCAGAAGCAACCGGGTGGAATATTGCTGGTGCAAC
AGT300
GGCAGGGCACAGTGCCACTCAGTGCCTGTCAAAAGTTGCAGCGAGCCAAGGTGTTT
CAAC360
GGGGGCACCTGCCAGCAGGCCCTGTACTTCTCAGATTTCGTGTGCCAGTGCCCCGAA
GGA420
TTTGCTGGGAAGTGCTGTGAAATAGATACCAGGGCCACGTGCTACGAGGACCAGGG
CATC480
AGCTACAGGGGCACGTGGAGCACAGCGGAGAGTGGCGCCGAGTGCACCAACTGGA
ACAGC540
AGCGCGTTGGCCCAGAAGCCCTACAGCGGGCGGAGGCCAGACGCCATCAGGCTGGG
CCTG600
GGGAACCACAACTACTGCAGAAACCCAGATCGAGACTCAAAGCCCTGGTGCTACGT
CTTT660
AAGGCGGGGAAGTACAGCTCAGAGTTCTGCAGCACCCCTGCCTGCTCTGAGGGAAA
CAGT720
GACTGCTACTTTGGGAATGGGTCAGCCTACCGTGGCACGCACAGCCTCACCGAGTCG
GGT780
GCCTCCTGCCTCCCGTGGAATTCCATGATCCTGATAGGCAAGGTTTACACAGCACAG
AAC840
CCCAGTGCCCAGGCACTGGGCCTGGGCAAACATAATTACTGCCGGAATCCTGATGG
GGAT900
GCCAAGCCCTGGTGCCACGTGCTGAAGAACCGCAGGCTGACGTGGGAGTACTGTGA
TGTG960
CCCTCCTGCTCCACCTGCGGCCTGAGACAGTACAGCCAGCCTCAGTTTCGCATCAAA
GGA1020
GGGCTCTTCGCCGACATCGCCTCCCACCCCTGGCAGGCTGCCATCTTTGCCAAGCAC
AGG1080
AGGTCGCCCGGAGAGCGGTTCCTGTGCGGGGGCATACTCATCAGCTCCTGCTGGATT

CTC1140
TCTGCCGCCCACTGCTTCCAGGAGAGGTTTCCGCCCCACCACCTGACGGTGATCTTG
GGC1200
AGAACATACCGGGTGGTCCCTGGCGAGGAGGAGCAGAAATTTGAAGTCGAAAAATA
CATT1260
GTCCATAAGGAATTCGATGATGACACTTACGACAATGACATTGCGCTGCTGCAGCTG
AAA1320
TCGGATTCGTCCCGCTGTGCCCAGGAGAGCAGCGTGGTCCGCACTGTGTGCCTTCCC
CCG1380
GCGGACCTGCAGCTGCCGGACTGGACGGAGTGTGAGCTCTCCGGCTACGGCAAGCA
TGAG1440
GCCTTGTCTCCTTTCTATTCGGAGCGGCTGAAGGAGGCTCATGTCAGACTGTACCCA
TCC1500
AGCCGCTGCACATCACAACATTTACTTAACAGAACAGTCACCGACAACATGCTGTGT
GCT1560
GGAGACACTCGGAGCGGCGGGCCCCAGGCAAACTTGCACGACGCCTGCCAGGGCGA
TTCG1620
GGAGGCCCCCTGGTGTGTCTGAACGATGGCCGCATGACTTTGGTGGGCATCATCAGC
TGG1680
GGCCTGGGCTGTGGACAGAAGGATGTCCCGGGTGTGTACACAAAGGTTACCAACTA
CCTA1740
GACTGGATTCGTGACAACATGCGACCGTGACCAGGAACACCCGACTCCTCAAAAGC
AAAT1800
GAGATCCCGCCTCTTCTTCTTCAGAAGACACTGCAAAGGCGCAGTGCTTCTCTACAG
ACT1860
TCTCCAGACCCACCACACCGCAGAAGCGGGACGAGACCCTACAGGAGAGGGAAGA
GTGCA1920
TTTTCCCAGATACTTCCCATTTTGGAAGTTTTCAGGACTTGGTCTGATTTCAGGATAC
TC1980
TGTCAGATGGGAAGACATGAATGCACACTAGCCTCTCCAGGAATGCCTCCTCCCTGG
GCA2040
GAAGTGGCCATGCCACCCTGTTTTCGCTAAAGCCCAACCTCCTGACCTGTCACCGTG
AGC2100
AGCTTTGGAAACAGGACCACAAAAATGAAAGCATGTCTCAATAGTAAAAGAAACAA
GAGA2160 TC2162 (2) INFORMATION FOR SEQ ID NO:4: (i) SEQUENCE
CHARACTERISTICS: (A) LENGTH: 16 base pairs (B) TYPE: nucleic acid (C)
STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi)
SEQUENCE DESCRIPTION: SEQ ID NO:4: SYUCARMGACUANANG16 (2)
INFORMATION FOR SEQ ID NO:5: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 4
base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii)
MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:5: GUCY4 (2) INFORMATION FOR SEQ ID
NO:6: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 6 base pairs (B) TYPE: nucleic
acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA
(xi) SEQUENCE DESCRIPTION: SEQ ID NO:6: TAGAGC6 (2) INFORMATION FOR SEQ

ID NO:7: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 25 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:7: GAGGACTTCAATAGTAGTATCCTGC25 (2) INFORMATION FOR SEQ ID NO:8: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 17 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:8: CAATTCGAGCTCGGTAC17 (2) INFORMATION FOR SEQ ID NO:9: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 9 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:9: AAAGGACTC9 (2) INFORMATION FOR SEQ ID NO:10: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 25 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:10: GTGATGGCCTGGCTCACCTGGACAA25 (2) INFORMATION FOR SEQ ID NO:11: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 5 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:11: TCTKC5 (2) INFORMATION FOR SEQ ID NO:12: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 4 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:12: TTTT4 (2) INFORMATION FOR SEQ ID NO:13: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 12 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:13: WSCWYSCTYCTG12 (2) INFORMATION FOR SEQ ID NO:14: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 11 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:14: HVGYKCASCTT11 (2) INFORMATION FOR SEQ ID NO:15: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 8 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:15: TYCTSWYC8 (2) INFORMATION FOR SEQ ID NO:16: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 28 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:16: CCAAGCTTAGACGAGGGGGAAAAGGGTT28 (2) INFORMATION FOR SEQ ID NO:17: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 27 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:17: CCAAGCTTGGAGGAGGGTGCCAGGGGG27 (2) INFORMATION FOR SEQ ID NO:18: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 27 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:18: CCAAGCTTGAAGCTCCTCAGAGGAGGG27 (2) INFORMATION FOR SEQ ID NO:19: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 28 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:19: CCAAGCTTTCATCAGATGGCGGGAAGAT28 (2) INFORMATION FOR SEQ ID NO:20: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 9 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA

(xi) SEQUENCE DESCRIPTION: SEQ ID NO:20: ARCTKGRTT9 (2) INFORMATION FOR SEQ ID NO:21: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 7 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:21: WTYCTCT7 (2) INFORMATION FOR SEQ ID NO:22: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 8 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:22: ARYCSGGG8 (2) INFORMATION FOR SEQ ID NO:23: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 7 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:23: GATAGAC7 (2) INFORMATION FOR SEQ ID NO:24: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 29 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:24: GGAAGCTTACTGGATGGTGGGAAGATGGA29 (2) INFORMATION FOR SEQ ID NO:25: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 2162 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:25:

TGAGCACAGGGCTGGAGAGAAAACCTCTGCGAGGAAAGGGAAGGAGCAAGCCGTG
AATTT60
AAGGGACGCTGTGAAGCAATCATGGATGCAATGAAGAGAGGGCTCTGCTGTGTGCT
GCTG120
CTGTGTGGAGCAGTCTTCGTTTCGCCCAGCCAGGAAATCCATGCCCGATTCAGAAGA
GGA180
GCCAGATCTTACCAAGTGATCTGCAGAGATGAAAAAACGCAGATGATATACCAGCA
ACAT240
CAGTCATGGCTGCGCCCTGTGCTCAGAAGCAACCGGGTGGAATATTGCTGGTGCAAC
AGT300
GGCAGGGCACAGTGCCACTCAGTGCCTGTCAAAAGTTGCAGCGAGCCAAGGTGTTT
CAAC360
GGGGGCACCTGCCAGCAGGCCCTGTACTTCTCAGATTTCGTGTGCCAGTGCCCCGAA
GGA420
TTTGCTGGGAAGTGCTGTGAAATAGATACCAGGGCCACGTGCTACGAGGACCAGGG
CATC480
AGCTACAGGGGCACGTGGAGCACAGCGGAGAGTGGCGCCGAGTGCACCAACTGGA
ACAGC540
AGCGCGTTGGCCCAGAAGCCCTACAGCGGGCGGAGGCCAGACGCCATCAGGCTGGG
CCTG600
GGGAACCACAACTACTGCAGAAACCCAGATCGAGACTCAAAGCCCTGGTGCTACGT
CTTT660
AAGGCGGGGAAGTACAGCTCAGAGTTCTGCAGCACCCCTGCCTGCTCTGAGGGAAA
CAGT720
GACTGCTACTTTGGGAATGGGTCAGCCTACCGTGGCACGCACAGCCTCACCGAGTCG
GGT780
GCCTCCTGCCTCCCGTGGAATTCCATGATCCTGATAGGCAAGGTTTACACAGCACAG
AAC840
CCCAGTGCCCAGGCACTGGGCCTGGGCAAACATAATTACTGCCGGAATCCTGATGG

GGAT900
GCCAAGCCCTGGTGCCACGTGCTGAAGAACCGCAGGCTGACGTGGGAGTACTGTGA
TGTG960
CCCTCCTGCTCCACCTGCGGCCTGAGACAGTACAGCCAGCCTCAGTTTCGCATCAAA
GGA1020
GGGCTCTTCGCCGACATCGCCTCCCACCCCTGGCAGGCTGCCATCTTTGCCAAGCAC
AGG1080
AGGTCGCCCGGAGAGCGGTTCCTGTGCGGGGGCATACTCATCAGCTCCTGCTGGATT
CTC1140
TCTGCCGCCCACTGCTTCCAGGAGAGGTTTCCGCCCCACCACCTGACGGTGATCTTG
GGC1200
AGAACATACCGGGTGGTCCCTGGCGAGGAGGAGCAGAAATTTGAAGTCGAAAAATA
CATT1260
GTCCATAAGGAATTCGATGATGACACTTACGACAATGACATTGCGCTGCTGCAGCTG
AAA1320
TCGGATTCGTCCCGCTGTGCCCAGGAGAGCAGCGTGGTCCGCACTGTGTGCCTTCCC
CCG1380
GCGGACCTGCAGCTGCCGGACTGGACGGAGTGTGAGCTCTCCGGCTACGGCAAGCA
TGAG1440
GCCTTGTCTCCTTTCTATTCGGAGCGGCTGAAGGAGGCTCATGTCAGACTGTACCCA
TCC1500
AGCCGCTGCACATCACAACATTTACTTAACAGAACAGTCACCGACAACATGCTGTGT
GCT1560
GGAGACACTCGGAGCGGCGGGCCCCAGGCAAACTTGCACGACGCCTGCCAGGGCGA
TTCG1620
GGAGGCCCCCTGGTGTGTCTGAACGATGGCCGCATGACTTTGGTGGGCATCATCAGC
TGG1680
GGCCTGGGCTGTGGACAGAAGGATGTCCCGGGTGTGTACACAAAGGTTACCAACTA
CCTA1740
GACTGGATTCGTGACAACATGCGACCGTGACCAGGAACACCCGACTCCTCAAAAGC
AAAT1800
GAGATCCCGCCTCTTCTTCTTCAGAAGACACTGCAAAGGCGCAGTGCTTCTCTACAG
ACT1860
TCTCCAGACCCACCACACCGCAGAAGCGGGACGAGACCCTACAGGAGAGGGAAGA
GTGCA1920
TTTTCCCAGATACTTCCCATTTTGGAAGTTTTCAGGACTTGGTCTGATTTCAGGATAC
TC1980
TGTCAGATGGGAAGACATGAATGCACACTAGCCTCTCCAGGAATGCCTCCTCCCTGG
GCA2040
GAAGTGGCCATGCCACCCTGTTTTCGCTAAAGCCCAACCTCCTGACCTGTCACCGTG
AGC2100
AGCTTTGGAAACAGGACCACAAAAATGAAAGCATGTCTCAATAGTAAAAGAAACAA
GAGA2160 TC2162 (2) INFORMATION FOR SEQ ID NO:26: (i) SEQUENCE
CHARACTERISTICS: (A) LENGTH: 13 base pairs (B) TYPE: nucleic acid (C)
STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi)
SEQUENCE DESCRIPTION: SEQ ID NO:26: GAAAACCTCTGCG13 (2) INFORMATION

FOR SEQ ID NO:27: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 30 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:27: ACGATGCATTCTGTAGAGAAGCACTGCGCC30 (2) INFORMATION FOR SEQ ID NO:28: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 13 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear (ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:28: CTCGAGCATTTTC13 (2) INFORMATION FOR SEQ ID NO:29: (i) SEQUENCE CHARACTERISTICS: (A) LENGTH: 15 base pairs (B) TYPE: nucleic acid (C) STRANDEDNESS: single (D) TOPOLOGY: linear

(ii) MOLECULE TYPE: cDNA (xi) SEQUENCE DESCRIPTION: SEQ ID NO:29: ATCGGGAWCCTGCGG15